

Requested Patent: JP11282636A

Title:

HIGH DENSITY RAID SUBSYSTEM WITH HIGHLY INTEGRATED CONTROLLER ;

Abstracted Patent: US6188571 ;

Publication Date: 2001-02-13 ;

Inventor(s):

BASCO JOSE PLATON (US); ROGANTI ADRIANO (US); WILLE THOMAS (US);
SMITH RONALD BRUCE (US) ;

Applicant(s): AIWA RAID TECHNOLOGY INC (US) ;

Application Number: US19970963841 19971103 ;

Priority Number(s): US19970963841 19971103 ;

IPC Classification: G06F1/16 ;

Equivalents: ;

ABSTRACT:

The present invention provides a method and apparatus for a mass storage subsystem such as a RAID array. The invention includes a housing which defines first and second cavities with the first cavity housing an array controller such as a RAID controller. The second cavity houses a plurality of substantially conventional IDE drives conforming to the 3.5" form factor. The array is configured to maximize cooling of the array controller and the drives within the extremely small space defined by the housing.

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平11-282636

(43) 公開日 平成11年(1999)10月15日

(51) Int. Cl. ⁶	識別記号	F I
G 0 6 F 3/06	5 4 0	G 0 6 F 3/06 5 4 0
G 1 1 B 33/02	3 0 1	G 1 1 B 33/02 3 0 1 F
33/14	5 0 3	33/14 5 0 3

審査請求 未請求 請求項の数 1 O L 外国語出願 (全 55 頁)

(21) 出願番号 特願平10-350665

(22) 出願日 平成10年(1998)11月4日

(31) 優先権主張番号 9 6 3 8 4 1

(32) 優先日 1997年11月3日

(33) 優先権主張国 米国 (US)

(71) 出願人 000000491

アイワ株式会社

東京都台東区池之端 1 丁目 2 番 11 号

(72) 発明者 アドリアーノ ロガンティ

アメリカ合衆国 フロリダ州 マーゲイ

ト, エヌ. ダブリュ. サーティファース

ト ストリート 5320

(72) 発明者 ロナルド ブレース スミス

アメリカ合衆国 フロリダ州 ウェリントン,

マイスティック ウェイ 1137

(74) 代理人 弁理士 浅村 皓 (外 3 名)

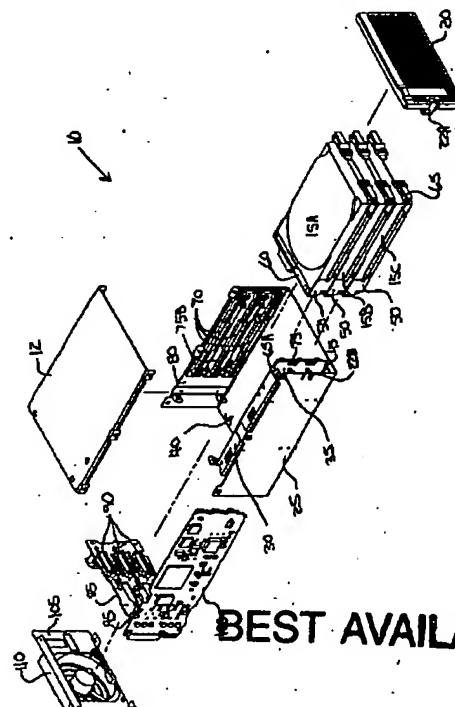
最終頁に続く

(54) 【発明の名称】 高集積度コントローラを有する高密度 RAID サブシステム

(57) 【要約】

【課題】 この発明は RAID アレイのような大容量記憶装置サブシステムのための方法と装置を提供する。

【解決手段】 この発明は、第1および第2の空洞を画定するハウジングを含み、前記第1空洞は、RAID コントローラのようなアレイコントローラを収容する。第2空洞は、3.5 インチ規格に適合する実質的に通常の IDE ドライブを複数個収納する。このアレイは、ハウジングにより限定される極端に小さな空間内で、アレイコントローラとドライブの冷却を最大にするように構成される。



BEST AVAILABLE COPY

【特許請求の範囲】

【請求項1】 ホストシステムへのSCSIインターフェイスと内蔵大容量記憶装置へのIDEインターフェイスを含む大容量記憶コントローラを収容する縦長の第1空洞と、

前記第1空洞の縦方向の軸が第2空洞の縦方向の軸に平行になるように、前記ホストシステムと通信可能な複数のIDEドライブを収容する前記第2空洞と、

そこを貫通するベントを有し、前記第2空洞内に収容されたIDEドライブの各々に接続するバックプレーンと、

前記第1空洞と前記第2空洞を通る充分な空気流を供給するために、前記バックプレーンとケースにより形成されるプレナムチャンバーを含む、大容量記憶アレイスブシステム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明はディスクドライブに関し、特にRAIDアレイスブシステムおよびコントローラに関する。

【0002】

【従来の技術】ハードディスク記憶装置は、全てのパーソナルコンピュータおよびサーバ、また同様に多くの他の関連するタイプのシステムに普及するようになった。多くの例において、そうした記憶装置は、少なくとも複数のバックアップとバックアップとの間の時間のためのミッションクリティカル(mission-critical)情報の保管のみを行う。結果としてこれらの記憶装置は信頼性が高く、極度に高度なデータ保全を維持しなければならない。

【0003】データ破壊を保証するために、多くのタイプの記憶装置サブシステムが開発されてきたが、その中にはミラーードライブ(mirrored drive)、フェイルオーバー(failover)システム、多重冗長ドライブサブシステムがある。その高い信頼性の故に特によく注目されてきた多重冗長サブシステムの一形式は、「安価なドライブの冗長アレイ」、すなわちRAIDサブシステムである。

【0004】一般にRAIDサブシステムはサーバその他のコンピュータシステム内に組み込まれてきた。一般に、RAIDサブシステムは二つまたはそれ以上のディスクドライブ(一般的に同一容量で、しばしば同一タイプ)を含み、そして少なくともRAID構成のいくつかの形式において、各ドライブがサブシステム上に記憶されるデータの最初の部分のための一次記憶装置として働くように構成され、またこのデータの第二の部分のためのバックアップ記憶装置として働くように構成される。RAIDシステムのための種々なバックアップスキームが開発されてきたが、その中にはRAID0、RAID1、RAID5がある。RAID0においてはデータの

冗長性が全く与えられず、RAIDアレイの容量は単純に個々のドライブの容量の合計である。RAID1においては、ミラーードライブ(mirrored drive)によく似た結合したドライブによってバックアップされる。RAID1はほとんどの場合偶数のドライブで構成される。これに対してRAID5は、典型的に最少三つのドライブから始まる種々な数のドライブにより構成される(二つのドライブは単にRAID1に程度が下がるだけである)。5ディスクRAID5サブシステムにとって、各ドライブはその容量の80%が一次記憶装置として働き、またその容量の20%が二次メモリとして働く。結果として、そうしたアレイの記憶容量はドライブの容量の合計の80%である。

【0005】一般に先行技術のRAIDサブシステムは、サーバの外部にあった。これはとりわけ、スペースと信頼性の問題を生じた。通常のサイズのPCケースは一般的にディスク記憶装置のために非常に限られた数のベイしか持たず、また通常のRAIDは利用可能なスペースに適合するにはあまりにも大きい。これはすでに混雑した領域内に特別な設置スペースを必要とし、また、RAIDデバイスへのサーバまたは他のPCを接続するための外部ケーブルのためのスペースも必要とする。より一般的な外部装置の故障の原因の一つはケーブル故障であり、これは、しばしば人間がケーブルに衝突したり不適切に切断したりすることによる。

【0006】いくつかの例において、例えばHPネットサーバラインのいくつかのモデルにおいて、記憶装置のために特別のベイを供給する特大のケースが供給されてきた。例えば、ネットサーバLM製品は、このサーバの拡張スロット内に挿入されるRAIDコントローラ付きの二倍幅ケースと、3.5インチ規格に合致するドライブのための八つのベイのスタックを有する。しかしながらこの解決方法は明らかに特定ベンダーの特定モデルのサーバを買うことを要しており、こうしてユーザのオプションを制限する。その上、このRAIDコントローラは、他の装置に使用できる拡張スロットを占有することとなる。先行技術のこれらの制約は、RAIDサブシステムを既存のサーバ内に含めるというユーザの希望を、非常に限られたオプションとしてきた。

【0007】この発明の譲受人は、既存サーバ内にRAIDサブシステムを含ませる際にエンドユーザが直面する二律背反のいくつかの面を解決することを、これまでに試みてきた。たとえばアイワ/コア(AIWA/core)のマイクロアレイは、5.25インチ標準高さ規格内に収まるように構成された。これによりこのサブシステムは、大部分の既存のケース内に設置可能になり、先行技術における置き場所(footprint)と外部接続の問題を避けることができた。このマイクロアレイ製品は、2.5インチ規格に合致する(最高5台までの)複数のIDEディスクドライブを、このサブシステ

ムへ挿入することを可能にする。このマイクロアレイ製品は、RAIDコントローラと、IDEドライブをこのRAIDコントローラにインタフェースし、ホストシステムへの外部SCSIインターフェースを供給するための関連電子製品とをその5.25インチ規格内に有する。

【0008】このマイクロアレイ製品は既存の先行技術に勝る多くの利点を提供するが、いくつかの欠点も有する。一つの重大な欠点はそれが高価な2.5インチディスクドライブを必要とすることであり、この2.5インチディスクドライブは、一般的に3.5インチ規格に合致するドライブよりもはるかに容量が少なく信頼性が低いもので、一方同時にかなりコストが高くなる。これらの制限の故に、2.5インチドライブは典型的にラップトップアプリケーションのみにおいて市場を見出したが、一方大部分のデスクトップアプリケーションは3.5インチドライブを使用してきた。

【0009】その上、マイクロアレイ製品のRAIDコントローラは、今日他の装置において利用できるものに比較して、制限されたスループットしか提供されておらず、複雑で従って高価な設計となっている。このコントローラは実質的に従来の考えをインプリメントし、またアレイの中の各ドライブごとに独立のI/Oチャネルを提供する。これにより大きなスペースを必要とし、2.5インチ規格よりも大きいドライブの使用を妨げている。

【0010】

【発明が解決しようとする課題】結果として5.25インチ標準高さのベイの従来のサーバケースに適合でき、同時にそのスペース内に統合されたコントローラを提供し、また低コストで大容量の3.5インチドライブの使用を可能にするRAIDサブシステムへの需要が存在してきた。

【0011】

【課題を解決するための手段】この発明は先行技術を大きく改良するRAIDサブシステムを記述するが、その改良点は、改良した容量、改良したスループット、より高い信頼性、より低いコストであって、また同時に単一の5.25インチ標準高さのベイの中にフィットするものである。この発明のRAIDサブシステムは、EIDEインタフェースを使用する複数の3.5インチディスクドライブの使用を含み、また同時に望ましい高速のデータ転送速度を有するホストシステムへのウルトラSCSIインターフェースを提供するものである。

【0012】前記の目的を達成するために、機械的および電気的なインターフェースの注意深い管理が、アレイ中の個別のドライブとコントローラの間に、またサブシステムとホストの間に必要とされてきたが、これは厳しく制限されたスペース内で所望の性能を得るためである。その上サブシステム内で通風のために利用できる空間が非常に限られているので、注意深い熱管理が必要と

されてきた。最後に前記要件が従来のコントローラ設計の使用を実質上不可能にしたので、この発明の一部として高度に集積したRAIDコントローラが開発された。この発明のコントローラは追加的な特徴として、この発明のRAIDサブシステムの機械的な設計の外側の領域に大きな利点を提供する。

【0013】上記の機械的、電気的、熱的、諸問題に加えて、この発明はエンドユーザによるメンテナンスが容易にできることを目的としており、サブシステム内に統合されたドライブヘンドユーザが容易にアクセス可能とする付加的な要件が加わる。これはエンドユーザがサブシステムのフロントパネルを取り除いて、一つまたはそれ以上のドライブを取り除くことを可能にする。これは、1997年9月16日に出願され、「ディスクドライブラッチ」と称し、この発明と同一の譲受人に譲渡され、本願に参考文献として組込まれた米国特許出願第08/931766号に記述された方法である。同時に、各ドライブの動作についての情報に関しエンドユーザは、少なくともこのサブシステムフロントパネルにステータスとアクセスの情報が配信されることを強く希望する。ユーザにそうした情報を提供する最も信頼できる方法は、RAIDコントローラが装着されるプリント基板にLEDまたは他のディスプレイデバイスを集積することであるが、そうした設計を行えば、エンドユーザが少なくともそのプリント基板の端に触れるかもしれない。この結果として、アクセス中にエンドユーザがサブシステムの内部に適当な注意をしない場合、大量の静電気放電すなわちESDから、コントローラ基板を保護しなければならない。

【0014】前記のようにこの発明のコントローラに対しては、先行技術に一般に見出されない多数の設計上の制約がある。これらの中には、空間の制限があり、それは制御基板のための規格内で利用できる空間が、単純に従来のコントローラの設計の使用を許容しないことである。第二に規格により課される熱の要件が、従来のコントローラの設計を過剰な熱を生ずるものとして受け入れ不可能にしている。第三にコストの要件が、複数のコントローラの使用を望ましくないものになっている。

【0015】結果として、高度に集積されたRAIDコントローラが開発され、そこではSCSIホスト機能により使用される単一のI/Oチャネルが供給され、またアレイ内に含まれる複数のドライブと共にDMA機能が供給される。この単一のI/Oチャネルは、時間多重化されていて、これにより各ドライブが所定の限られた時間にコントローラへアクセスできるようになっており、またインタフェースのSCSIホスト部分が、同様に所定の限られた時間にコントローラへアクセスできるようになっている。適当なクロック速度を使用することにより、この単一チップコントローラは、それに必要な諸機能の各々に対応すると同時に、必要なDMA機能を管理

することができる。一つの実施例において、コントローラのエンジンは、在庫品のフィールドプログラムマブルゲートアレイ、すなわちFPGAで実施できるが、その設計はASICまたは他の類似の装置によっても実施できる。ここではこの発明のコントローラを内蔵RAIDサブシステムに使用する例を示したが、内蔵および外部の両方のRAIDサブシステムへの適用、または全く外部のRAID環境への適用にまでこの設計はおよぶ。

【0016】その上、この発明のアレイは、アレイ内に維持されるディスクドライブのホットスワッピングを可能にする。ユーザがアクセス可能なドライブ特定スイッチにより、このシステムのファームウェアはドライブをパワーダウン可能である。それからこのドライブは除去されて、新たなドライブが設置される。それからファームウェアは自動的に新ドライブの設置を検知して、電力を再び印加すると共に、データと制御信号を再接続する。この技法は、ダウンタイムやデータの損失なしに遂行できるメンテナンスを可能にし、パワーサージを抑制し、静電気放電からの保護を行う。

【0017】この発明の、これらおよび他の特徴は、以下の「発明の実施の形態」を添付図面と共に参照することにより、一層良く理解されるであろう。

【0018】

【発明の実施の形態】全体に図1ないし図8、特に図1と図2を参照すると、この発明のRAIDサブシステムが、より良く理解できる。後でより良く理解されるように、このサブシステムのトップカバー12は、図1では除去されているが、図2に明示されている。複数の従来型IDEコンプライアントディスクドライブ15A、同15B、同15C（IDEはその一般的な範囲にEIDEとウルトラDMAドライブを含む）は、その各々が受け入れられた3.5インチ規格に従っており、ケース25内のフロントベゼル（bezel）20の後ろに装着されている。ケース25は、ベゼル20と共働して、ほぼ幅5.25インチ、高さ3.25インチとして一般に受け入れられている、5.25インチ規格の標準の高さに一致する。ラッチ22Aは、ベゼル20と一体に形成され、ケース25内で受け皿22Bと噛み合わされているが、このラッチ22Aは、ケース25の内側に係合するベゼルの反対端の内側のL型ポスト（図示なし）と共働して、ベゼルがラッチから外れて、外へ開いてメンテナンスのために、取りはずされるようにする。規格の長さの制約はもっと緩いが、一般に8インチないし10インチ程度である。内部の上板30と内部の側壁35が、ケース25に固定されて、3.5インチドライブ15Aないし15Cを装着するに適した第1空洞40を画定する。上板30と側壁35はまた、第1空洞40の左に、長く狭い第2空洞45を囲むが、これについては後で詳述する。

【0019】各ドライブ15Aないし同15Cは、U型

ドライブブラケット50（図8と図9に最も良く示されこれに関連して詳細に説明されている）内に装着され、U型ドライブバスケット50は一組のレール55A-Bおよびドライブ拡張ボード60を含んでなる。装着機構65がレール55A-B上に装着され、この機構は1997年9月16日に出願され「ディスクドライブラッチ」と題し本願に参考文献として組み入れられた米国特許出願第08/931766号に記述されている。レール55A-Bは、装着プレート75A-B（図2と図4に最も良く出ている）に合わせて複数の溝70内に、滑動可能にはめ込まれており、装着プレート70A-Bは、ケース25の右側壁80の内側と内部側壁35の右面に取り付けられている。

【0020】ドライブ15Aないし同15Cの各々に結合されたドライブ拡張ボード60の後ろに位置しているのは、バックプレーン85で、図6と図7に関連して以下に説明する。バックプレーン85は、各ドライブ拡張ボード上でマッピングコネクタ90Aと噛み合わせるために、複数のコネクタ90（図2と図7に特に示される）を有し、またケース25の左側の下の空洞45内に装着されるRAIDコントローラプリント基板100を装着するために、コネクタ95（図6と図7に最も良く出ている）を有する。バックプレーン85とRAIDコントローラ基板100を収納し、ファン110を支持するために、後ろカバープレート105がケース25の後ろに取り付けられている。後ろカバープレート105は、バックプレーン85の後ろに間隔をとって、アルミチャンバー115を形成し、これによりケース25により窮屈に配置されたRAIDコントローラ基板100および複数のドライブ15A-Cをファンが効率的に冷却できるようにしている。上記の種々の素子の他の詳細は、他の図面に関連して以下に説明する。

【0021】続けて全体に図1ないし図8を参照し、また特に図3を参照すると、ディスクドライブ15A-Cの配置とバックプレーン85に対するそれらの接続がより良く理解できる。ドライブ15A-C（ドライブ15Aのみが図3に示されている）はラッチ機構65によりケース25の中にラッチされ、ラッチ機構65はドライブ拡張基板60に取り付けられたコネクタ90Aをバックプレーン85上のコネクタ90と噛み合い接続させる。ドライブ拡張ボードがドライブ15Aの後ろにいくらか間隔をとっておかれ、これにより特に複数のドライブ15A-Cの長さのバラツキを許容し、また気流チェンバも形成している。同様にコネクタ90と同90Aの空間もドライブ拡張ボード60とバックプレーン85の間に気流チェンバ150を形成する。ドライブ15Aが、フレキシブルリボンケーブル60Aにより、ドライブ拡張ボードに接続されているのがここに見えるが、図9には一層よく見える。リボンケーブル60Aは、ドライブ15Aに含まれるIDEコネクタに接続し、また異

なったタイプのドライブ上のコネクタの位置の僅かな変動を許容する。

【0022】バックプレーン85は、(左で)装着ブラケットの上部および下部ベア155と、もう一つのベア160により、ケース25に取り付けられる。例示的な実施例において、内側側壁35と一体に形成された装着ブラケット155は、二重に曲げられている。装着ブラケット160は、側壁80へ取り付けられている。多くの場合必要ないが、装着ブラケット155内のこの二重の曲げにより提供される弾性は、挿入および除去の処理によりドライブとバックプレーンに加えられる歪みの力を吸収するのを補助する。更に、装着ブラケットとバックプレーンの弾性は、リボンケーブル60Aと共に、ファンまたは他のドライブまたはシステム内の他のどこからかにより加えられる全ての振動から、ドライブを分離するのを補助すると考えられる。この組み合わせは、システムの信頼性を増し、ドライブの寿命を延ばすのを補助すると考えられる。少なくとも、いくつかの場合において、バックプレーン85とドライブ拡張ボード60の柔軟性は、リボンケーブル60Aと共に、適切な柔軟性を与えると共に分離を行うのに充分である。

【0023】プレナムチャンバー115もまた図3から理解され、ファン110の正面の減圧空間に見ることができる。プレナムチャンバー115は、空洞40を通じてドライブ15A-Cの周りに引き込まれ、空洞150内に集められた空気と共に、空洞45を通じてRAIDコントローラ基板100を通過して引き込まれた空気を集める。バックプレーン85と後部カバープレート105の間の間隔は、ファン110がRAIDアレイを通じて空気を引き込み、許容可能な温度範囲内にアレイを維持する効率を最適化するように必要なだけ調節できる。

【0024】製造を容易にするために、RAIDコントローラ基板は、空洞45内に滑動可能に装着される。2組のガイド165は、上部壁30に下向きに穴をあけて実質的にスロットを形成することにより、単一的に形成可能であり、ケース25の底に形成された類似のスロット(図示なし)と組み合わせて、基板100の上縁を空洞45内の中心に位置させる。同様のガイド170もまた壁30の正面に設けられる。

【0025】図4を参照すると、ドライブ15A-Cをスタックする配列が、空洞40と45を通る空気流と同様により良く理解できる。図3では、上部カバーが示されていない。RAIDコントローラ基板100は、空洞45の中央に配置され、基板のいずれの側からも空気流が通れる。その上、装着ブロック75A-Bとレール70の間隔が、空洞40内で、ドライブ15A-Cのいずれの側にも空気を通過させる。空洞40と45を通る空気流に一致するように、ファン110とプレナムチャンバーのサイズを適当に定めることにより、これらのドライブとRAIDコントローラ基板は充分な冷却がな

されて、長期連続運転を可能にする。より薄いドライブを使用して、図11に関連して説明するRAIDコントローラに相応な変更を行えば、ドライブを追加し得る。

【0026】更に、リーフスプリング175が、空洞45の正面に配置され、それにより基板100を正しい位置へ付勢すると共に、アレイのメンテナンス中にユーザにより基板に印加されうる静電気電荷を放電すべく、基板100のアースも行う。コンピュータシステム内の大部分のサブシステムと異なり、RAIDコントローラ100の前縁は、ベゼルをはずすだけで、コンピュータシステムのフロントパネルからユーザによりアクセス可能となる。結果として、ESD用の接地に適当なパスが、基板100の1方側の少なくとも一部分を導電物質でメッキして、そのメッキをリーフスプリング165を通じてケースに接続することにより提供される。このリーフスプリングは、一般的に、銅または他の適当なスプリング材料で形成し得る。少なくともいくつかの場合に、そうした他の材料の銅メッキが望ましい。基板100のメッキは図10Aに最も良く示され、参照番号285でメッキが識別される。

【0027】更に図4に示されるのは、各ドライブの1組のLED18A-Bと、各ドライブの押しボタン185である。LED185Aは典型的に、関連するドライブのステータスを示し、異なった動作状態を示すのに異なった色を用いる多色LEDである。LED180Bは一般的に、関連するドライブの活動を示す。押しボタン185は、RAIDコントローラに、ユーザによる関連ドライブの切断指示信号を与える。押しボタン185を押すことにより、RAIDコントローラは関連ドライブへの電力と信号バスを切断して、残りのアレイを動作し続けながら、そのドライブを安全に取りはずしできるようにする。ドライブが一旦アレイから電気的に切断されると、ラッチ65により、このドライブが物理的に取りはずし可能状態となる。このドライブまたは他の同等なドライブが、それからラッチ65を締めることによりアレイへ戻される。例示的な実施例において、バックプレーンコネクタ90へドライブコネクタ90Aを挿入すると、アレイによりドライブの付加が検知される。しかしながら、いくつかの実施例においては、再び関連の押しボタンを押すことにより、新しいドライブの追加をアレイが検知するようにしてもよい。

【0028】次に図5を参照すると、このサブシステムの後部は上面半横向き透視図に見られ、冷却ファン110とホストを接続するための外部コネクタを、より良く理解できるようになっている。図3に関連して説明したように、冷却ファン110はバックプレーンの中央の背後に配置され、ファンからの受け入れがたい乱気流を十分に避け得るように間隔をあけてあり、ドライブとプリント基板を通る空気流の量を増加させ、従ってファンの効果を最大化している。ファン110の左に、9ピンDシ

エルコネクタ200が配置され、これはこの発明の譲り受け人により提供されるアレイビュー (Array View) 製品またはサブシステムのステータスをモニターするのに適した他の製品のようなモニタ装置に接続されて使用される。Dシェルコネクタ200の下には、通常のパワーコネクタ205がある。D型コネクタ200とパワーコネクタ205は、例示的な実施例において、バックプレーン85に接続され、後部カバープレート105の開口を通じて延びている。ファンの右側に、シングルエンド超ワイドSCSI標準に合致した高密度コネクタ210があり、これと共に、ユニットのIDの設定、種々の診断および他の通常の機能の遂行のためにふさわしいジャンパブロック215がある。コネクタ210と215は、RAIDコントローラ基板100に取り付けられ、後部カバープレート105の開口を通して延びている。SCSIコネクタ210は、ホストシステムへのインターフェイスを提供し、またこのサブシステム全体が、ホストシステム内のホストアダプタに対して単一のSCSI装置に見える。他の実施例において、このサブシステムは、ディファレンシャルSCSI、ワイドSCSI、または他のインターフェイスなどの異なったコネクタの異なったインターフェイス規格に準拠するようにして良い。

【0029】次に図6を参照すると、後ろカバープレート (ファンを含む) を除去した状態のアレイサブシステムの後立面図が示され、それによりバックプレーン85のレイアウトが詳細に示されている。また図7を参照すると、前立面図にバックプレーン85のレイアウトを示す。特に図6を参照すると、コネクタ200と205がバックプレーン85と一体に示され、デュアルコネクタ95がバックプレーン85をRAIDコントローラ基板100に接続する方法も示されている。その上、多様なベントすなわちカットアウト225が、バックプレーン85の周辺と内部の両方に存在して、プレナム (plenum) チャンバー115への空気流を改善することが分かる。バックプレーンは孔230を貫通する4つのネジ (図示なし) により保持され、装着ブラケットの噛み合うペア155と160へ装着される。またバックプレーン85の上にアラーム235があり、これはアレイのパフォーマンスをモニタする多様なセンサからの信号に応答する。例えば、これは、1つまたはそれ以上のドライブ温度センサ240、ファンセンサ245などを含み、実施例では、図7に示すように、バックプレーンの正面に装着される。更に、バックプレーンの正面に示されるコネクタ90は、高サイクル低挿入力コネクタで、通常のIDEバスとパワーの両方を、関連のドライブに供給する。それからドライブエクステンション60はドライブへ適当な機械的インターフェイスを供給するが、これには通常のIDEコネクタと通常のパワーコネクタによる。コネクタ90へ接続するドライブ15A-Cの

特有の順序は重要でないが、本例示的な実施例では、トップコネクタに結合されるドライブは、ドライブ0と指定され、ミドルコネクタはドライブ1と指定され、ボトムコネクタはドライブ2と指定される。

【0030】次に図8と図9を参照すると、単一のドライブ15Aをドライブブラケット50にはめ込む方法がより良く理解できる。図1と図2に関連して前記したように、ドライブブラケット50は一組のレール55A-Bと共にドライブ拡張ボード60を有してなる。ドライブ15は普通の機械ねじによりブラケット50に取り付けられ、またケーブル60Aおよびバックプレーンコネクタ90と共に通常のAmphenolパワーコネクタ60Bを通じてブラケットへ電気的に接続されている。またラッチ機構65が示される。

【0031】次に図10と図11を参照すると、RAIDコントローラ基板100が見られる。RAIDコントローラ基板100は単一の両面プリント回路基板で、それは図12からより良く理解される。図10Aに示される側はアウトボード (outboard) 側であるが、こちらから見るとコネクタ210と同215は左端に見られる。RAIDコントローラはRAIDエンジン集積回路260 (これはフィールドプログラムマブルゲートアレイ、ASIC、または他の適当な実装であってもよい) を含み、これにより必要な待ち行列とDMAの機能を遂行する。RAIDエンジン260は、キャッシュメモリ265 (図11)、RAIDコントローラの動作を管理するためのRISC CPU270、それに結合したCPUメモリ275 (両方とも図11)、ホストインタフェースを管理するためのSCSIプロセッサ280 (図10) と通信する。LED180A-Bと押しボタン185が、前縁でRAIDコントローラボードに接続されるのが見え (図11)、一方ボードの反対側に、導体ESDメッキ285 (図3に関連して一般に議論される) が見える。本願に示されるRAIDコントローラボード100の例示的な実施例は、バックプレーンをボード100に接続できるようにするために一組のコネクタ275を含む。時間/日付チップ290と共に、図12に関連して記述した通常の機能を遂行する種々の他のセンサと論理回路も供給され得る。メッキ285の配列から見て、特に図10から分かることは、この発明のサブシステムのメンテナンスを行うユーザが、メンテナンス中に、静電気を放電して伝達し、結果としてRAIDコントローラを損傷するのを実質的に防止することになっていることであり、それはメッキ285が上記のように接地面に直接接続されているからである。

【0032】次に図12および図13ないし図16を参照すると、この発明の電氣的動作がよりよく理解される。一般に、RAIDサブシステムはホストから見ると、通常のSCSIコマンドに外部的に適合する単一ボリュームに見えるが、しかし内部的には完全にRAID

アレイとして動作する。このRAIDアレイの動作はRAIDコントローラにより制御され、一方RAIDコントローラは、時分割多重と独立の32ビットDMAとCPUソフトウェア処理メモリを使用することにより、エンジンのピーク速度において同時非競合アクティビティを可能にする。DMAすなわちキャッシュメモリ265は、たとえば1×36メモリとして構成された4メガバイトであり、160MB/秒程度の帯域幅を有するシングルサイクルページ化EDOパイプラインを供給する。CPUメモリ275は、例示的な実施例では1×32で形成された4メガバイトとして構成され、80MB/秒の帯域幅を有する2サイクルページ化EDOパイプラインを提供する。

【0033】CPU270は、たとえば40MHzで動作するLSI LR33310-40 32ビットRISCプロセッサであり、埋め込みRAIDオペレーティングシステムを記憶するFLASH ROM300と共働する。アーキテクチャの中心にRAID集積回路260があり、これはたとえばAlteraフィールドプログラマブルゲートアレイ(FPGA)またはその均等物またはASICとして構成され、これは各DMA I/Oチャネルのためにコマンド待ち行列の提供、種々のデータI/O待ち行列の管理、それに関連するキーバス上のバスアクティビティの管理を行い、システムの周辺機能をサポートする。FPGA260に関連する5つの主要なバスは次の通りである：40MB/秒の16ビットSCSIプロセッサバス305（典型的にウルトラSCSI動作のために構成されているが他のSCSIプロトコルをサポートできる）；3.33MB/秒の8ビットSCSIチップパイプライン化I/Oバス310；16MB/秒の16ビットIDEドライブバス315；160MB/秒の36ビットディスクキャッシュメモリ(DCM)バス320；および80MB/秒の32ビットCPUバス325である。FPGA260は、5つのバス全てが並列に動作できるように構成され、SCSIプロセッサバス305、IDバス315、およびDCMバス325によりRAIDエンジン260へのアクセスを多重化するのに十分な速度で動作するRAIDエンジン260を有し、また定義されたサイクル内でSCSIプロセッサバス305とIDEバス315の各々にワンタイムスロットをDCMバス320のためにツータイムスロットを割り当てることにより、IDEバス315と、DCMバス320を有する。例示的な実施例において、1サイクル全体で100ns程度であり、4つのタイムスロットは各々25nsを割り当てられる。RAIDエンジン260のパフォーマンスの故に、このサブシステムの正味スループットは、主として以下の4つの要素に依存する：IDEドライブのパフォーマンス、一体型オペレーティングシステム内のRAID機能オーバーヘッド、ユーザのホストアダプタのパフォーマンス、ユーザ

ホストアプリケーションのドライバオーバーヘッドである。

【0034】更に図12を参照すると、システムの動作は実質的に次の通りである：電源投入すると、フラッシュROM300からRISCプロセッサ270に関連するCPUメモリ275へ、オペレーティングシステムをロードすることにより、システムは安定状態になる。初期化の後、ある時点で、読み取りまたは書き込みの要求を、ホストシステム340からホストSCSIバス350で受け取り、これが適当でアレイば、終端ブロック355により終結される。この要求はそれからSCSIプロセッサ280により処理され、SCSIプロセッサ280は適当な信号をRAIDエンジン260へI/Oバス310上で送り、適当な確認信号を送り返される。それからこのデータは、ホストシステムによりSCSI DMAバス305上で利用可能にされる。この時点で、ディスクキャッシュメモリ265は空である。この要求が情報を書き込むことでアレイば、CPU270はRAIDエンジン260に、このデータをDCM265へ渡すように命令し、DCM265でキャッシュ中に保持される。その後、バックグラウンド処理の間に、バス320上でRAIDエンジン260により最初にデータをアクセスすることにより、（使用されるRAID stripingに従って、）適当にディスクに割り当てられ、IDEバス上で、ISOディスクバッファ360A-Cへ書き出される。それからこのデータは、特定のディスク15A-Cに書き出される。本願に記述する例示的なシステムにおいては、バス320が10本のアドレス線と36本のデータ線を含むのが適当である。同様に、バス325は、32本のアドレス線と32本のデータ線を有している。この処理は、RAIDエンジンからSCSIプロセッサ280へ供給され、それからホスト340へ供給される確認信号で終了する。これらの種々の事象のタイミングは、図13ないし図16および図17に関連して一層詳細に説明する。図13において、DPIとDPOはそれぞれ「入力データバス」と「出力データバス」を表す。

【0035】読み取り動作においては、この処理は実質上類似しているがいくらか逆になっている。この処理はSCSIインタフェースをアクティブにし、一般的には開始の時に行われるようにする。それからホストは確認/肯定応答信号を送り設定を実行し、これに続いてPIOバス310上で特定データの要求を送る。それからこの要求はRAIDエンジン260により検出され、RAIDエンジン260はそれをドライブ15へ渡す。このデータはこれらのドライブからRAIDエンジン260へ返され、ここからそれは一時的な記憶のためにディスクキャッシュメモリ265へ渡される。適当な時間にCPU270は、このデータがバス320を介してDCM265から読み出されて、エンジン260を通じてES

EIプロセッサ280へデータバス305上で渡されるようにする。それからこのデータはESEIプロセッサ280からホストへバス350上で渡される。

【0036】RAIDエンジンはそのデータ管理機能の他に、多数の周辺機器補修機能を管理する。これらの中には温度超過検出器240とファンエラー検出器245が含まれ、アラーム235で(適当な時に)警告信号を生成する。時間/日付クロック290もまたモニタされ、その電力は、システムがオフであるときに電力は電池または他の電源365により供給される。ハードウェア検出線370はステータスレジスタ375によりモニタされる。ドライブのための電力サージ制御はバッファ380でモニタされる。このサブシステムの監視はまたduart385上でも供給される。典型的な監視は監視と保守の両方のためにRS232リンク390上で遂行される。

【0037】次に図13ないし図16および図17を参照するとRAIDエンジン260の動作の詳細が一層良く理解され、その中にはRAIDエンジンにより種々なバス上の信号が多重化されるタイミングが含まれる。上記の図と同様に、図12から同じ部品は同じ参照番号を割り当てられている。

【0038】前と同様に、ホストシステムの電源が投入されると、RAIDサブシステム10が初期化を行い、フラッシュROM300内に維持されるソフトウェアにより設定されるときにRISC CPU270により一連のイネープリングファクターが生成される。これらのイネープリングファクターは、IDEドライブを知られた状態にし、またSCSIプロセッサ280をアクティブでイネーブルされた状態にし、ホストシステム340へ通知する。ホストシステムはSCSIプロセッサからの通知を確認し肯定的応答をする。その上、このイネープリングファクターは、RAIDエンジン260を知られた状態にし、また特にRAIDエンジン260の内部にある40MHz I/Oコマンド待ち行列プロセッサを初期化する。図15は、図14の鎖線内の一部分の代案の配列を示す。

【0039】初期化の後に、ホストシステムは前と同様にドライブ15A-Cに書き込むべきデータを送る。この情報はヘッダ情報とデータを含んでなり、バス350上でSCSIプロセッサ280へ供給される。SCSIプロセッサ280で加工した後に、このヘッダ情報は図13に破線で示したRAIDエンジン260へ、8ビットプログラマブルI/Oバス310で供給される。このヘッダ情報はRAIDエンジン260へSCSI PIO400を通して送られるが、SCSI PIO400は入力データバス405と出力データバス410を有する。入力データバス405はマルチプレクサmux415の片側に接続し、マルチプレクサmux415は間接的にI/Oコマンド待ち行列プロセッサ390へ入力を

供給する。このI/Oコマンド待ち行列プロセッサ390はフレームベースのスクリプトプロセッサであって、ハーフワードのコマンド、行アドレスコマンドおよび列アドレスコマンドを、16ビットバスを経由してレジスタ395へ供給する。レジスタ395はまたRISCプロセッサ270からアドレスを受け取る;プロセッサ270はまた、muxバッファ399を通じて与えられるバス325の10ビットブランチ397を経由して、DCM265へアドレスを供給する。レジスタ395の出力は、パイプラインレジスタ401とバッファ403を通じて、DCM265へ多重化10ビットバス(1メガバイトのアドレス空間をアドレスする)を経由して供給され得る。レジスタ395の出力は、間接的に、RAIDエンジン260に関連する他の場所を示す出力データバスを供給するが、これにはmux415への第2の入力が含まれる。

【0040】SCSIプロセッサ280からのデータが供給されるヘッダ情報と並行して、SCSIプロセッサ280からSCSI PIO400へバス310上をRAIDエンジン260へ16ビットDMAバス305経由で供給される。特にデータは40MHzで作動する16ビットツ-32ビットファネル420へ与えられるが、これはRAIDエンジン260が32ビット幅で内部的に動作するからである。データはファネルmux425の片側へ供給されて、それからI/O待ち行列mux430の片側へ供給される。待ち行列mux430の出力はフレームベースI/O待ち行列435へされるが、フレームベースI/O待ち行列435は40MHzで作動し256x32で構成されて、160MB/秒のスループットを供給する。I/O待ち行列435への他の入力、は、種々のIDEプリンタ440、SCSIポインタ445、DCMポインタ450を有する。データはI/O待ち行列435の出力へ全部クロックされた第1パイプライン出力レジスタ455へ供給され、それからDCMmux460の片側へ供給される。muxの出力は第2パイプライン出力レジスタ465へ供給され、バッファ470を経由して、それからRAIDエンジン260からDCM265へ供給される。

【0041】ディスク15A-Cの適便な一つに書き込むのに適当になるまで、データはDCM265に記憶され、書き込むディスクは典型的に通常のアルゴリズムに従って、RAIDオペレーティングシステムにより決定される。その時に、I/Oコマンド待ち行列390が、ディスクドライブへデータを書き込むコマンドを発行する。データはDCM265によりバッファ475へ供給され、それからパイプライン入力レジスタ480へ供給される。データはそれから第2入力レジスタ485と共に、プロセッサ入力mux490の片側へ供給される。ドライブを書き込むために、データがレジスタ485を通じてmux430の他の側へ与えられ、それからI/

O待ち行列435へ与えられる。

【0042】I/O待ち行列435からのデータはSCSI I/Oパネル420へ供給され、またディスクI/Oファネル500へも供給される。ディスクI/Oファネル500は出力データを、ディスクドライブ15との通信のために、32ビットデータ幅から16ビットデータ幅へ再変換する。ディスクドライブとの通信についてのその他のことは、図12に関連して説明した通りである。

【0043】ホストシステム340によりRAIDサブシステムに要求されるもう一つの典型的な動作はRAIDサブシステムからのデータ検索である。データ検索はホストシステム340から開始され、ここでもホストの要求をPIOバス310、SCSI PIO400を介してRAIDエンジン260へ供給し、それからmux415を通じてI/Oコマンドプロセッサ390への入力データバス405へ供給する。I/Oコマンドプロセッサ390は、それから適当なRAC/CACアドレスをレジスタ395経由で供給し、これによりデータが検索されるようにする。

【0044】ホストシステムにより希望されるデータの適当なアドレスが、DCM265へ供給される。もしデータがDCM265内に維持されれば、それはレジスタ480と同485を経由してmux430へ供給され、それからI/O待ち行列435へ供給される。I/O待ち行列435からこのデータがSCSI I/Oファネル420へ供給され、ここで出力データが16ビット幅へ変換される。それからデータはDMA305上をSCSIプロセッサ280へ供給され、最終的にホスト340へバス350上を出力される。

【0045】しかしながらホストにより要求されたデータがキャッシュ265内に現在維持されていなければ、このデータをディスクに要求しなければならない。この場合は、要求されたデータのアドレスがレジスタ480、485およびmux430を経由して、I/O待ち行列435へ供給される。それからI/Oの待ち行列435の出力が、ディスクI/Oファネル500へ供給されてドライブ15A-Cへ出力される。それからデータは、必要な待ち時間の後にドライブから検索され、その後にドライブから到着するデータが、ディスクファネル500内で16ビット幅から32ビット幅へ変換される。ファネル500のデータの出力はそれからmux425の第2の側へ供給され、そこからI/O待ち行列435へ第2のmux430を通じて供給される。

【0046】それからI/O待ち行列435の出力は、I/Oファネル420を通じて出力データについて上記したのと同じ方法で供給され、その結果このデータは通常の方法でホストシステムへ供給される。フィールドプログラマブルゲートアレイ(FPGA)に基づくインプリメンテーションのために、FPGAルートROM49

2が供給され、電源投入に際してFPGAがパーソナリ化されるようにする。ASICまたは他のゲートアレイのインプリメンテーションにおいては、そうしたルートROMは不必要である。同様にRAID OSがCPUメモリ275へ電源投入に際してロードされ、全てのソフトウェア制御がCPUメモリ内に記憶された命令から引き出される。

【0047】この発明の動作の重要な特徴は、単一のRAIDエンジン260がSCSIプロセッサ、ディスク、プロセッサ270と通信するのに必要な多重DMA待ち行列を管理できることである。この目的はデータをRAIDエンジン260へ供給するキーバスを時間多重化することにより達成される。これは40MB/秒以上では作動しない他の装置に比較して、このI/O待ち行列が160MB/秒の効果的な速度で動作する故に可能である。これによりRAIDエンジンはその約4分の1の時間をSCSIプロセッサとディスクドライブの各々へ割り当てることができ、またその約半分の時間をDMAアドレッシングに割り当てることができる。図17に示すタイミング図は、この発明の実施例のいくつかの面に重要な時分割多重化を提供するのに必要な位相化アドレッシングを与えるものである。

【0048】特にI/O待ち行列435とスクリプトプロセッサの40MHzクロックが600で示され、一方RAIDへのアクセスするためのSCSI位相が605で示されている。IDE位相は610で示され、一方DCM位相は615で示される。付加的な機能において、IDEドライブからアクセス要求されない時にサイクルが起こった場合は、DCMによる使用のために位相が再割り当てされる。同様にSCSIプロセッサがI/Oアクセス要求しないサイクルについてはSCSIプロセッサに割り当てられる位相がDCMへ再割り当てされる。こうしてこの発明により極度に高いスループットが達成できることが理解されよう。

【0049】次に図18を参照すると、この発明のホットスワッピング装置、(アレイの残りの部分が動作し続ける間に一つまたはそれ以上のドライブを除去できる)がより良く理解できる。特に、例えばあるドライブの故障のために、ユーザがドライブ15A-Cの一つを取りはずすことを希望する場合、取りはずすべきドライブに関連する押しボタンスイッチ185をユーザが作動させる。これがCPU270に信号を与えて、ソフトウェア制御のもとに作動するCPU270がFPGA制御論理回路へ信号を与えて、関連するドライブの12ボルトと5ボルトの電源700と705をそれぞれ切断させる。その上プロセッサによりデータバス701と制御バス715が、サブシステムの他の部分から電気的に切断される。この時点でユーザはラッチ65をやり直し、必要なドライブを除去する準備ができている。

【0050】ドライブを再設定するには、ユーザはドラ

イブをドライブベイに挿入しラッチ65をラッチすることにより、処理の機械的部分を逆にたどるだけでよい。双安定ラッチがドライブの再挿入を検知して、新しく挿入されたドライブへの電力と信号の接続を再び加えるように、CPU270へ信号を与える。この方法で、古いドライブが除去されて、新ドライブが設定され得る。

【0051】従って、RAIDのための新しく新規なサブシステムと、高度に統合されたコントローラが記述されたことが、理解されたであろう。当業者にとっては、これらの教示が与えられれば、開示された発明を組み込んだ多数の代案や均等物が存在することが理解されよう。結果として、この発明は、先行する例示的な実施例により限定されるべきものではなく、むしろ請求項によって限定されるべきである。

【図面の簡単な説明】

【図1】上部カバーを除去したこの発明のRAIDサブシステムの正面半横向き透視図を示す。

【図2】この発明のサブシステムの種々の構成要素の半横向き分解透視図を示す。

【図3】上部カバーを除去したこの発明のRAIDサブシステムの上面図を示す。

【図4】前カバーを除去したこの発明のRAIDサブシステムの前立面図を示す。

【図5】上部カバーを除去したこの発明のサブシステムの後ろ半横向き透視図を示す。

【図6】後ろカバープレート除去したこの発明のサブシステムの後ろ立面図を示し、特にバックプレーンを示す。

【図7】この発明のバックプレーンのレイアウトを、前立面図で示す。

【図8】単一のドライブと関連の装着ブラケットを、バックプレーンインターフェイス基板と共に、上面半横向

き透視図で示す。

【図9】図8の装着ブラケットの透視図を示し、特に、ドライブとブラケットの間のリボンケーブルインターフェイスを示す。

【図10】RAIDコントローラ基板の片側を、レイアウト形式で示す。

【図11】RAIDコントローラ基板の第2の側を、レイアウト形式で示す。

【図12】RAIDエンジンを含むこの発明のRAIDコントローラを、概略ブロック図形式で示す。

【図13】図12のRAIDエンジンの内部構成を示す。

【図14】図12のRAIDエンジンの内部構成を示す。

【図15】図12のRAIDエンジンの内部構成を示す。

【図16】図12のRAIDエンジンの内部構成を示す。

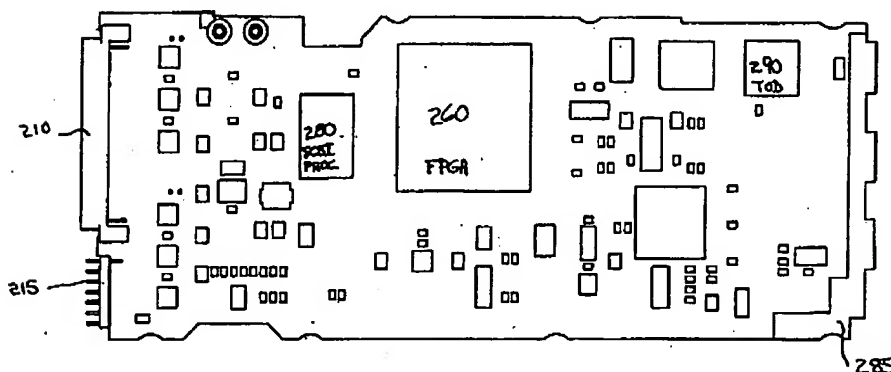
【図17】図12に示すRAIDエンジンの種々の動作のタイミングを示す。

【図18】RAIDアレイのホットスワップ機能を、略図形式で示す。

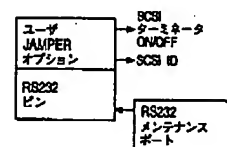
【符号の説明】

- 10 RAIDサブシステム
- 15A、15B、15C IDEドライブ
- 25 ケース
- 40 第1空洞
- 45 第2空洞
- 85 バックプレーン
- 100 RAIDコントローラ基板
- 115 プレナムチャンバー

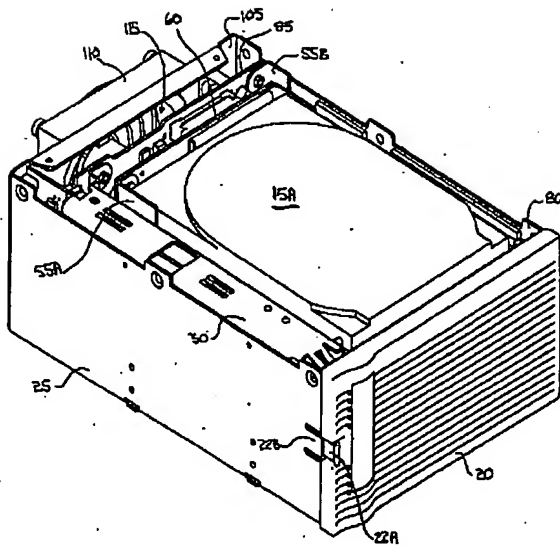
【図10】



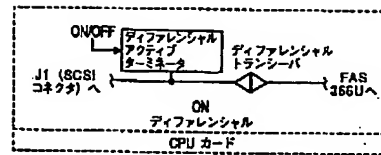
【図16】



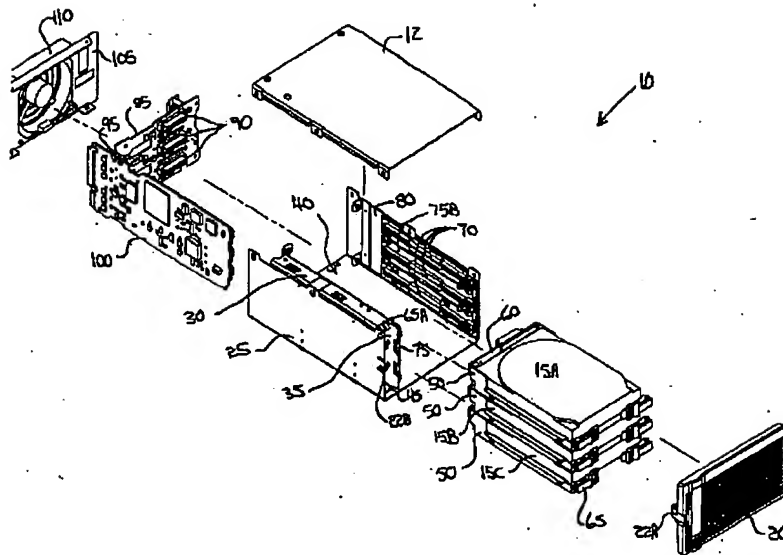
【図1】



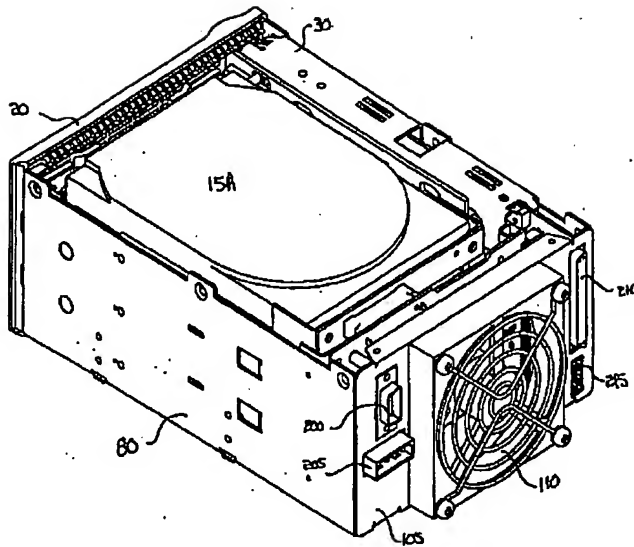
【図15】



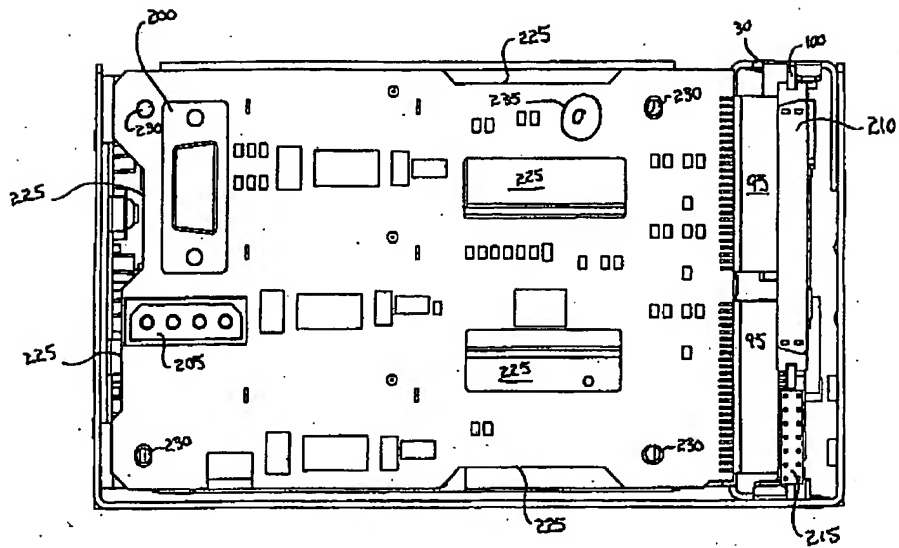
【図2】



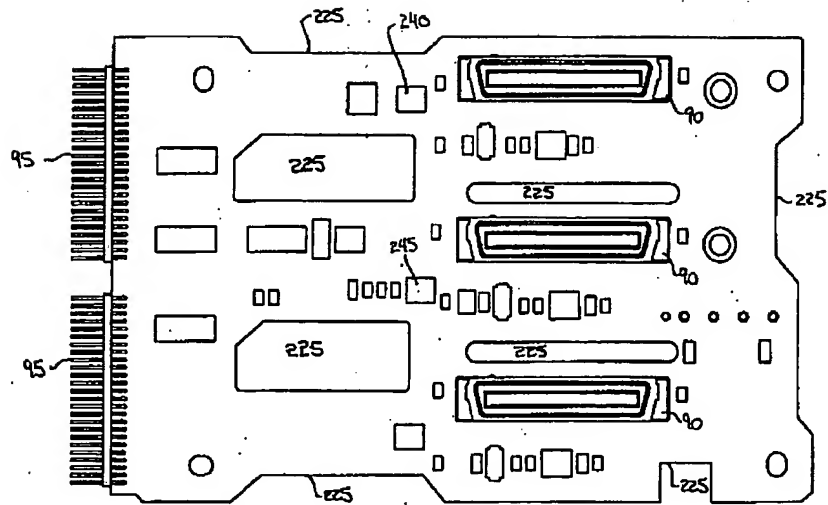
【図5】



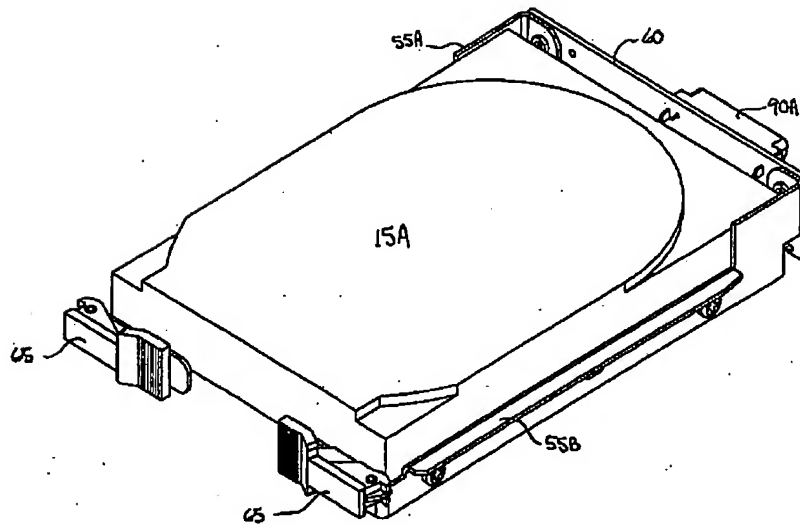
【図6】



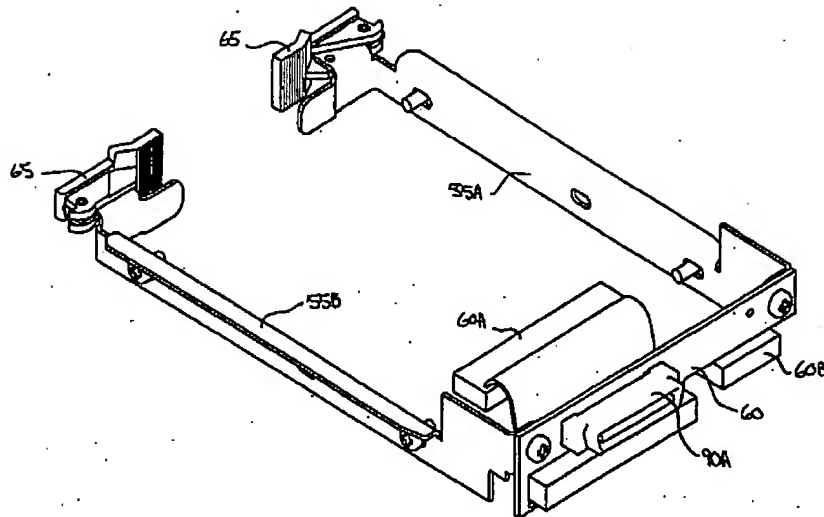
【図7】



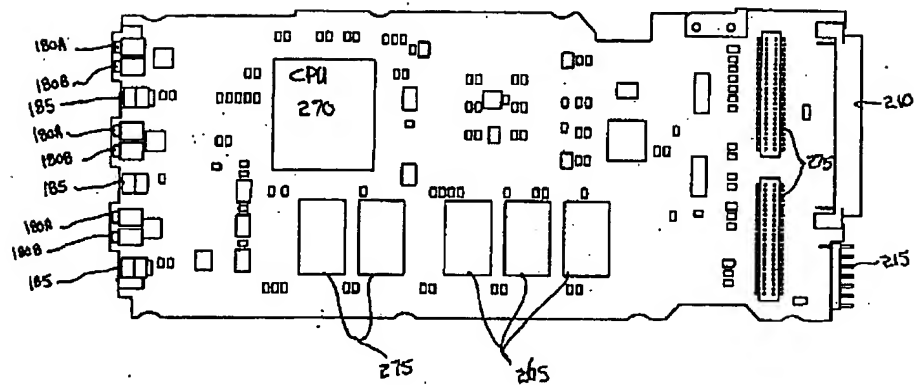
【図8】



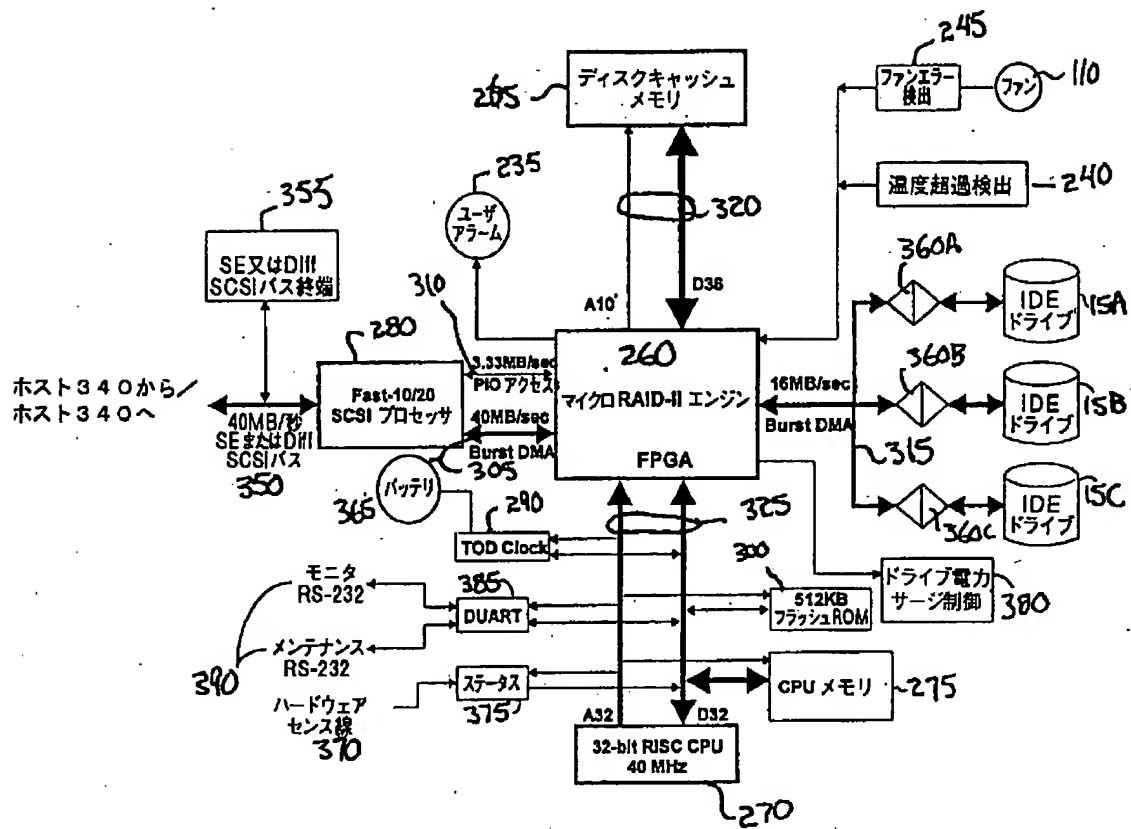
【図9】



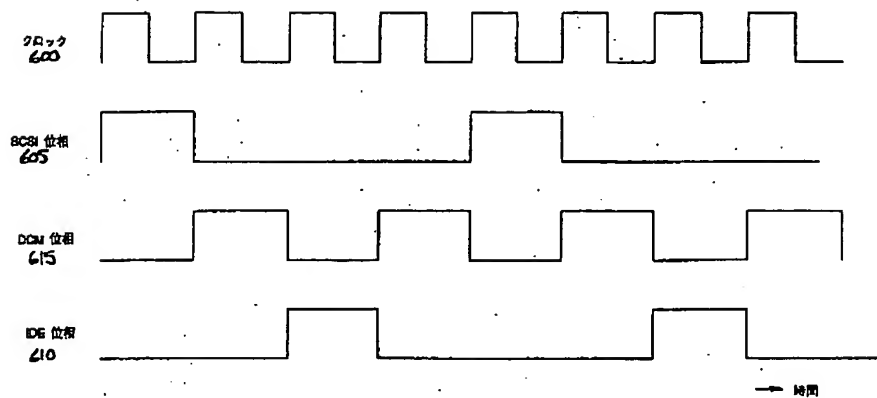
【図11】



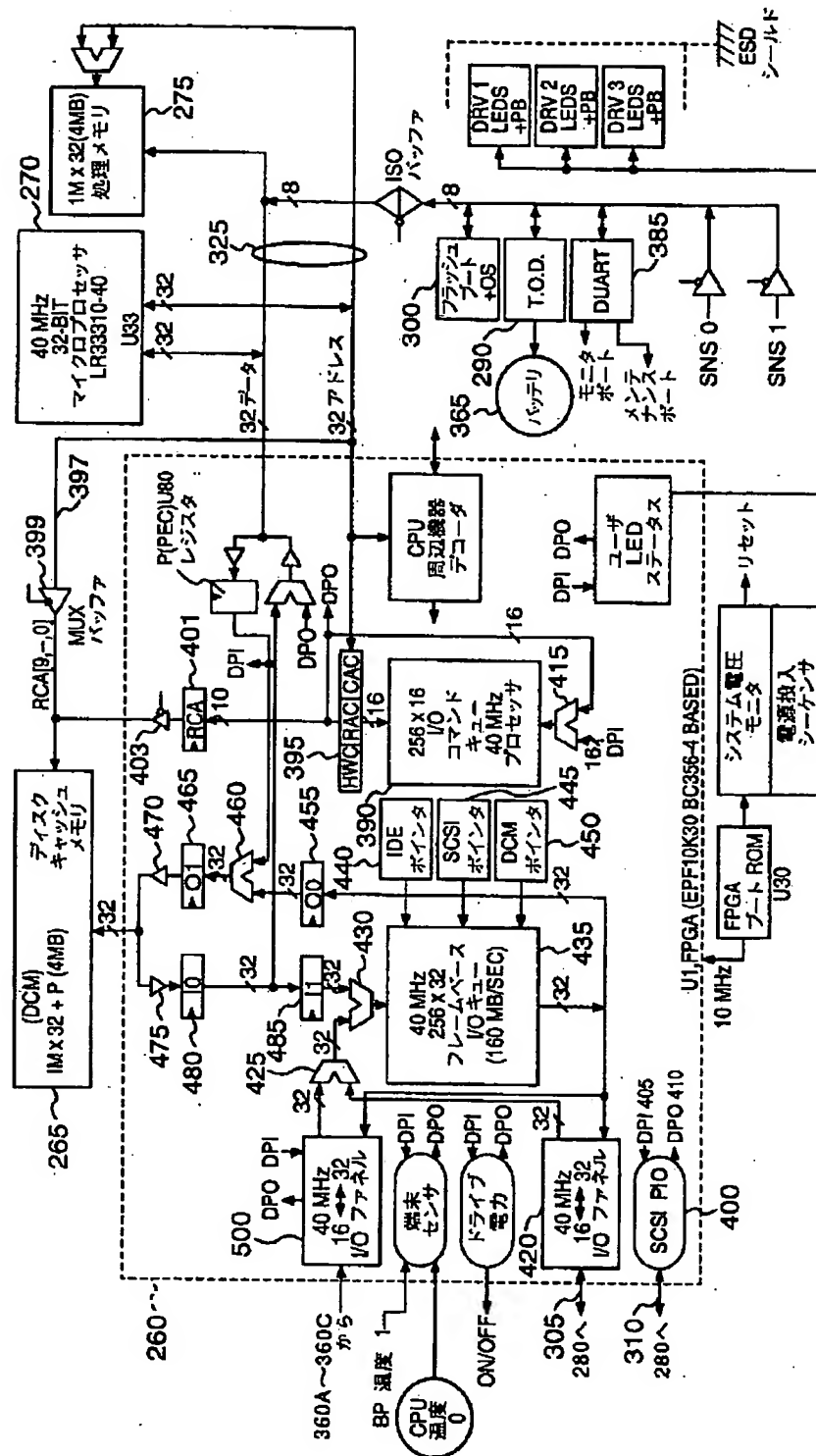
【図12】



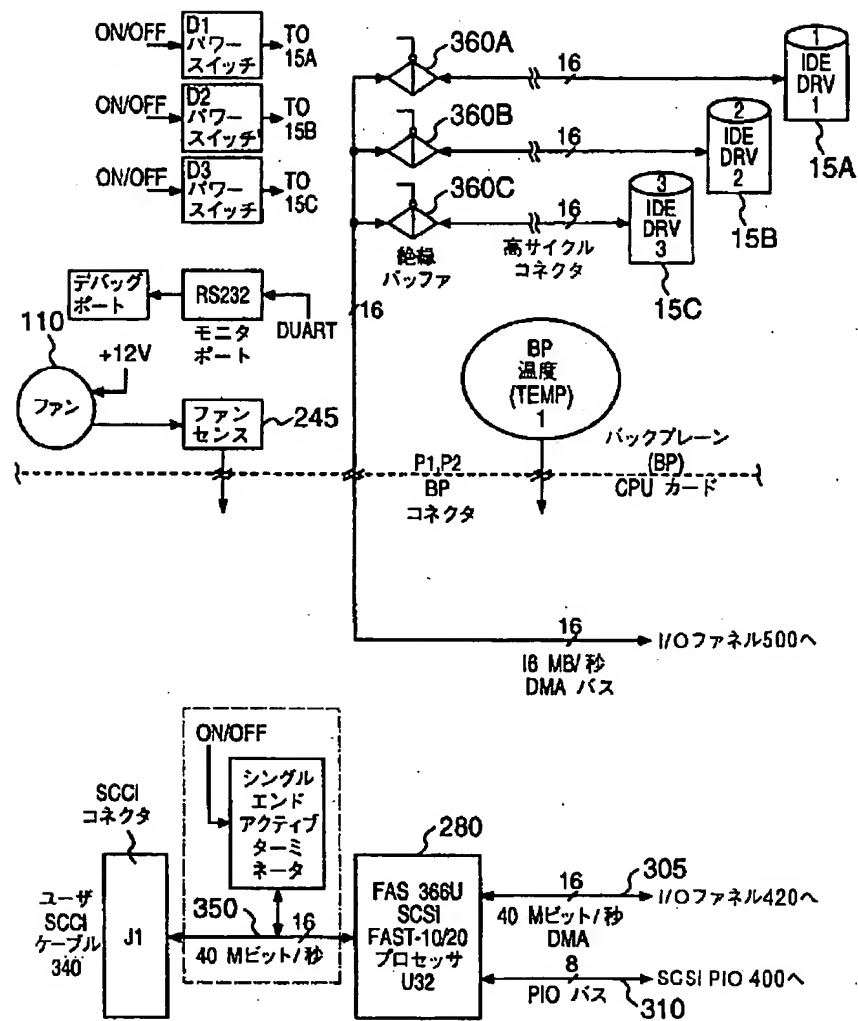
【図17】



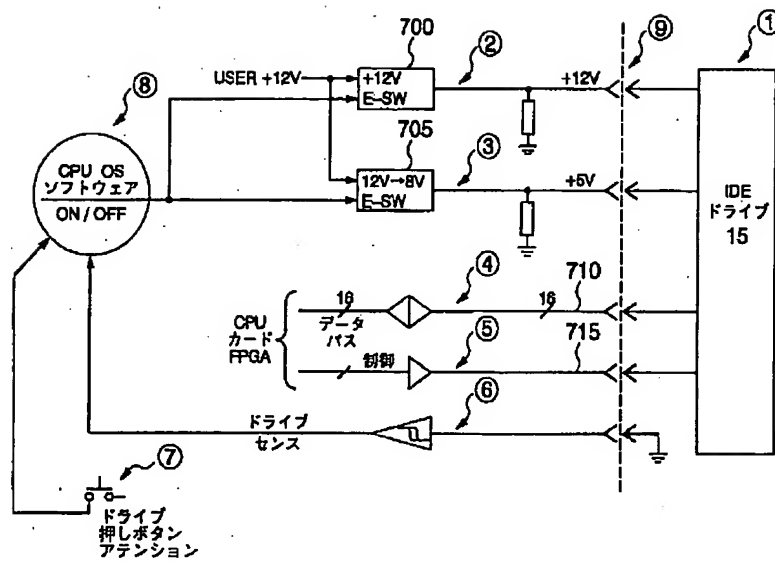
【図13】



【図14】



【図18】



フロントページの続き

(72)発明者 ホセ プラトン バスコ
 アメリカ合衆国 フロリダ州 ウェリントン,
 ニアンティック テラス 1276

(72)発明者 トマス ウイリイ
 アメリカ合衆国 フロリダ州 レイクワース,
 シダー ハースト コート 7627

【外国語明細書】

EXPRESS MAILING LABEL CM303714700US

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICATION

5

FOR

UNITED STATES PATENT

10

FOR

HIGH DENSITY RAID SUBSYSTEM
WITH HIGHLY INTEGRATED CONTROLLER

15

Inventors:

Adriano Roganti

Thomas Wille

Ron Smith

Jose Basco

20

SPECIFICATIONFIELD OF THE INVENTION

25 The invention relates to disk drives, and more particularly to RAID array subsystems and controllers.

BACKGROUND OF THE INVENTION

Hard disk storage has become ubiquitous for virtually every personal computer
30 and server, as well as many other related types of systems. In many instances, such storage represents the only repository for mission-critical information for at least the time between backups. As a result, these storage devices must be highly reliable and maintain extremely high data integrity.

Many types of storage subsystems have been developed to ensure against
35 data corruption, including mirrored drives, failover systems, and multiply redundant drive subsystems. A form of multiply redundant subsystem which has become particularly well-regarded for its high reliability is the "redundant array of inexpensive

EXPRESS MAILING LABEL EM303714786US

2

drives, " or RAID subsystem.

RAID subsystems typically have been implemented in servers and other computer systems. In general, RAID subsystems include two or more disk drives (typically of the same capacity, and frequently of the same type) and, in at least some forms of RAID implementations, are configured such that each drive serves as the primary storage device for a first portion of the data stored on the subsystem and serves as the backup storage device for a second portion of the data. Various backup schemes for RAID systems have been developed, including RAID 0, RAID 1, and RAID 5. In RAID 0, no data redundancy is provided, and the capacity of the RAID array is simply the sum of the capacities of the individual drives. In RAID 1, each drive is backed up by an associated drive much like mirrored drives. RAID 1 is implemented in most instances with even numbers of drives. RAID 5, on the other hand, can be implemented by a varying number of drives, typically beginning at a minimum of three (two drives would simply degrade to RAID1.) For a five-disk RAID 5 subsystem, each drive serves as primary storage for 80% of its capacity, and secondary storage for 20% of its capacity. As a result, the storage capacity of such an array is 80% of the sum of the capacities of the drives.

In general, prior art RAID subsystems have been external to the server. This has imposed space and reliability issues, among other things. Conventional sizes of PC cases typically offer only a very limited number of bays for disk storage, and conventional RAID arrays are simply too large to fit the available space. This imposes the requirement for extra floor space in what is typically already a crowded area, but also imposes the requirement for an external cable to connect the server or other PC to the RAID device. One of the more common causes of failures for external devices is cable failure, often due to human error in bumping or inadvertently disconnecting the cable.

In some instances, for example some models of the HP NetServer line, oversized cases have been provided which provide extra bays for storage devices. For example, the NetServer LM product includes a double-wide case with a RAID controller inserted into an expansion slot of the server and a stack of eight bays for drives conforming to the 3.5" form factor. However, this solution obviously requires buying a specific vendors specific model of server and thus limits the user's options. Moreover, the RAID controller occupies an expansion slot which might otherwise be available for other devices. These constraints of the prior art have left those wishing to include RAID subsystems in their existing servers with very limited options.

The assignee of the present invention has previously attempted to resolve some aspects of the dilemma presented to end-users attempting to include RAID

EXPRESS MAILING LABEL EM303714786US

3

subsystems in their existing servers. For example, Aiwa/Core's MicroArray is a RAID subsystem configured to fit within the 5.25" full height form factor. This permits the subsystem to be installed within most existing cases and therefore avoids the footprint and external connection issues of other prior art. The MicroArray product
5 permits a plurality of IDE disk drives (up to five) conforming to the 2.5" form factor to be inserted into the subsystem. The MicoArray product includes within its 5.25" form factor the RAID controller and related electronics necessary to interface the IDE drives to the RAID controller and to provide an external SCSI interface to the host system.

10 While the MicroArray product offered many advantages over existing prior art, it did have some drawbacks. One significant drawback was that it required the use of expensive 2.5" disk drives, which typically offer far less capacity and less reliability than drives conforming to the 3.5" form factor while at the same time costing significantly more. Because of these limitations, 2.5" drives have typically found a
15 market only in laptop applications, while most desktop applications have used 3.5" drives.

In addition, the RAID controller of the MicroArray product offered limited throughput compared to that available in other devices today and comprised a complicated — and therefore expensive — design. The controller implemented
20 substantially conventional wisdom and offered independent I/O channels for each of the drives in the array. This imposed significant space requirements which prevented the use of any drive larger than those complying with the 2.5" form factor.

As a result there has been a need for a RAID subsystem which is capable of fitting with a 5.25" full height bay of a conventional server case, while at the same
25 time offering an integrated controller within that space and the use of low-cost, high capacity 3.5" drives.

SUMMARY OF THE INVENTION

30 The present invention describes a RAID subsystem which substantially improves upon the prior art in offering substantially improved capacities, improved throughput, higher reliability, and lower cost while still fitting within a single 5.25" full height bay. The RAID subsystem of the present invention includes the use of a plurality of 3.5" disk drives using the EIDE interface, while at the same time offering
35 the Ultra-SCSI interface to the host system with its desirable high-speed data transfer rate.

To achieve the foregoing, careful management of the mechanical and

EXPRESS MAILING LABEL EM303714786US

4

electrical interfaces has been required — both between the individual drives in the array and the controller, and between the subsystem and the host — to fit the desired capabilities within a tightly limited space. In addition, careful thermal management has been required because of the very limited availability of space for airflow within 5 the subsystem. Finally, the foregoing requirements substantially prohibit the use of conventional controller designs, such that a highly integrated RAID controller has been developed as part of the present invention. The controller of the present invention has the additional feature of offering substantial benefits in areas outside the mechanical design of the present RAID subsystem.

10 In addition to the mechanical, electrical and thermal problems described above, the present invention is intended to permit ease of maintenance by the end user, which imposes the additional requirement of permitting the end user to have easy access to the drives integrated into the subsystem. This has been achieved by permitting the end-user to remove the front panel of the subsystem, which allows the 15 end-user to remove one or more of the drives in the manner described in U.S. Patent Application S.N. 08/931,766, filed on 9/16/97 and entitled Disk Drive Latch, assigned to the assignee of the present invention and incorporated herein by reference. At the same time, the end-user's desire for information on the operation of each drive substantially demands that status and access information be delivered to at least the 20 front panel of the subsystem. While the most reliable method for providing such information to the user is by integrating LEDs or other display devices into the printed circuit board on which the RAID controller is mounted, implementing such a design also imposes the limitation that the end user may also be able to touch at least an edge of that printed circuit board. This results in the requirement that the controller 25 board be protected from significant amounts of electrostatic discharge, or ESD, in the event the end-user does not take adequate precautions while accessing the interior of the subsystem.

As noted previously, the controller of the present invention is subject to multiple design constraints not generally found within the prior art. Included in these 30 are space limitations, in that the space available within the form factor for the controller board simply does not permit the use of conventional controller designs. Second, the thermal requirements imposed by the form factor reinforce that conventional controller designs are unacceptable as generating excessive heat. Third, cost requirements make the use of multiple controllers undesirable.

35 As a result, a highly integrated RAID controller has been developed in which a single I/O channel is provided for use by the SCSI host functions and the drives included within the array, as well as for DMA functions. The single I/O channel is

EXPRESS MAILING LABEL EM303714786US

5

time-multiplexed to permit each drive to access the controller for a prespecified, finite period, and also to permit the SCSI host portion of the interface to access the controller for a similar prespecified finite period. By the use of suitable clocking rates, the single-chip controller can thus attend to each of its required functions while at the
5 same time managing the requisite DMA functions. In one embodiment, the engine of the controller may be implemented in an off-the-shelf field programmable gate array, or FPGA, although the design may also be implemented in an ASIC or other similar device. While the controller of the present invention is shown herein used with internal RAID subsystems, the design has application for both internal and
10 external RAID subsystems and may also have application entirely outside the RAID environment.

In addition, the array of the present invention permits hot-swapping of disk drives maintained within the array. Activation of a drive-specific switch accessible to the user causes the firmware of the system to power down the drive. The drive may
15 then be removed and a replacement drive installed. The firmware then automatically senses the installation of the new drive, and reapplies power as well as reconnecting data and control signals. The technique allows maintenance to be performed without down time or loss of data, suppresses power surges and provides protection from electrostatic discharges.

20 These and other features of the present invention will be better appreciated from the following Detailed Description of the Invention, taken in conjunction with the attached Figures.

BRIEF DESCRIPTION OF THE DRAWINGS

25

Figure 1 shows in front three-quarter perspective view the RAID subsystem of the present invention with the top cover removed.

Figure 2 shows in exploded front three-quarter perspective view the various components of the subsystem of the present invention.

30 Figure 3 shows in top plan view the RAID subsystem of the present invention with the top cover removed.

Figure 4 shows a front elevational view of the RAID subsystem of the present invention with the front cover removed.

Figure 5 shows a rear three-quarter perspective view of the subsystem of the
35 present invention with the top cover removed.

Figure 6 shows a rear elevational view of the subsystem of the present invention with the rear cover plate removed, and in particular shows the backplane.

EXPRESS MAILING LABEL EM303714786US

6

Figure 7 shows in front elevational view the layout of the backplane of the present invention.

Figure 8 shows in top three-quarter perspective view a single drive and associated mounting bracket with backplane interface board.

5 Figure 9 shows in perspective view the drive mounting bracket of Figure 8, and in particular shows the ribbon cable interface between the drive and the bracket.

Figure 10~~A~~ shows in layout form one side of the RAID controller board.

Figure 10~~B~~¹¹ shows in layout form the second side of the RAID controller board.

Figure 10¹²~~A~~ shows in schematic block diagram form the RAID controller of the present invention including the RAID engine.

Figures 10^{13 to 16}~~A~~ shows the internal configuration of the RAID engine of Figure 10¹².

Figure 10¹⁷~~A~~ shows the timing of various operations of the RAID engine shown in Figure 10¹².

Figure 10¹⁸~~A~~ shows in schematic form the hot-swap capability of the RAID array.

15

DETAILED DESCRIPTION OF THE INVENTION

Referring generally to Figures 1 through 8, and particularly to Figures 1 and 2, the RAID subsystem 10 of the present invention can be better appreciated. As will be appreciated better hereinafter, the top cover 12 of the subsystem has been removed in Figure 1 but is evident in Figure 2. A plurality of conventional IDE compliant disk drives 15A, 15B and 15C (where IDE includes within its general scope EIDE and Ultra DMA drives), each of which also complies with the accepted 3.5" form factor, are mounted behind a front bezel 20 and within a case 25. The case 25 cooperates with the bezel 20 to fit within the conventional 5.25" full height form factor, which is generally accepted as approximately 5.25" wide by 3.25" high. A latch 22A, formed integrally with the bezel 20 and mated to a receiver 22B in the case 25, cooperates with L-shaped posts (not shown) on the inside of the opposite end of the bezel which engage the inside of the case 25 to permit the bezel to be unlatched, swung out and removed for maintenance. The length of the form factor is less tightly controlled but is generally on the order of eight to ten inches. An internal top plate 30 and internal side wall 35 are rigidly affixed to the case 25 to define a first cavity 40 suitable for mounting the 3.5" drives 15A-C. The top plate 30 and side wall 35 also enclose a second, long, narrow cavity 45 to the left of the first cavity 40, the use of which is discussed in greater detail hereinbelow.

Each drive 15A-C is mounted within a U-shaped drive bracket 50 (best seen in Figures 8 and 9 and described in detail in connection therewith) which comprises

EXPRESS MAILING LABEL EM303714785US

7

a pair of rails 55A-B and a drive extension board 60. A mounting mechanism 65 is mounted on the rails 55A-B, which mechanism is better described in U.S. Patent Application S.N. 08/931,766, filed on 9/16/97 and entitled Disk Drive Latch, referred to hereinabove and incorporated herein by reference. The rails 55A-B slidably fit 5 within grooves 70 in matching mounting plates 75A-B (best seen in Figures 2 and 4), which are affixed to the interior of the right sidewall 80 of the case 25 and the right face of the internal sidewall 35.

Positioned behind the drive extension boards 60 associated with each of the drives 15A-C is a backplane 85, described hereinafter in connection with Figures 6 10 and 7. The backplane 85 includes a plurality of connectors 90 (shown particularly in Figure 2 and Figure 7) to mate with a matching connector 90A on each of the drive extension boards, and also includes a connector 95 (best seen in Figures 6 & 7) for mounting to a RAID controller printed circuit board 100 mounted within the cavity 45 down the left side of the case 25. A rear cover plate 105 is affixed to the rear of the 15 case 25 to enclose the backplane 85 and the back edge of the RAID controller board 100, and supports a fan 110. The rear cover plate 105 is spaced behind the backplane 85 to form a plenum chamber 115 to permit the fan to cool efficiently the RAID controller board 100 and the drives 15A-C in the tight spacing imposed by the case 25. Other details of the various elements mentioned above will be described 20 in connection with other Figures.

Still referring generally to Figures 1-8 and with reference particularly to Figure 3, the arrangement of the disk drives 15A-C and their connection to the backplane 85 can be better appreciated. The drives 15A-C (only drive 15A is shown in Figure 3) are latched into the case 25 by virtue of latching mechanism 65, which 25 urges the connector 90A affixed to the drive extension board 60 into mating contact with the connector 90 on the backplane 85. It will be appreciated that the drive extension board 60 is spaced somewhat behind the drive 15A to permit, among other things, variations in the length of the drives 15A-C and also to provide an airflow chamber. Likewise, the spacing of the connectors 90 and 90A creates an airflow 30 chamber 150 between the drive extension board 60 and the backplane 85. The drive 15A can be seen to be connected to the drive extension board by a flexible ribbon cable 60A, visible here but better seen in Figure 9. The ribbon cable 60A connects to the IDE connector included with the drive 15A, and allows for slight variations in the location of the connector on different types of drives.

35 The backplane 85 is affixed to the case 25 by virtue of an upper and lower pair of mounting brackets 155 (at the left) and another pair 160. The mounting brackets 155, which are, in an exemplary embodiment, integrally formed with the internal side

EXPRESS MAILING LABEL EM303714786US

8

wall 35, may be seen to be double bent. Mounting brackets 160 may be seen to be affixed to the sidewall 80. While not necessary in many cases, the additional resiliency offered by the double bend in mounting bracket 155 aids in absorbing the deflection forces imposed on the drive and the backplane by the insertion and
5 removal process. In addition, the resiliency of the mounting brackets and the backplane, as well as the ribbon cable 60A, are believed helpful in isolating the drives from any vibration imposed by the fan, the remaining drives or elsewhere in the system. The combination is believed helpful in increasing the reliability of the system and extending the life of the drives. In at least some instances, the flexibility of the
10 backplane 85 and the drive extension board 60, together with the ribbon cable 60A, will be sufficient to provide adequate resiliency and isolation.

The plenum chamber 115 may also be appreciated from Figure 3, and can be seen to form a decompression space in front of the fan 110. The plenum chamber 115 collects air drawn around the drives 15A-C through cavity 40 and collected in
15 cavity 150 as well as air drawn past the RAID controller board 100 through cavity 45. The spacing between the backplane 85 and rear cover plate 105 can be adjusted as necessary to optimize the efficiency of the fan 110 in drawing air through the RAID array and maintaining the array within an acceptable thermal range.

For ease of manufacturing, the RAID controller board is slidably mounted
20 within the cavity 45. Two pairs of guides 165, which may be formed unitarily with the top wall 30 by being punched downward essentially to form a slot, position the top edge of the board 100 centrally within the cavity 45, in combination with a similar slot (not shown) formed in the bottom of the case 25. A similar guide 170 may also be provided at the front of the wall 30.

25 Referring next to Figure 4, the stacking arrangement of the drives 15A-C can be better appreciated as well as the airflow through the cavities 40 and 45. As with Figure 3, the top cover is not shown. The RAID controller board 100 can be seen to be centrally located in the cavity 45, permitting airflow past either side of the board 100. In addition, the gaps between the mounting blocks 75A-B and the rails 70 can
30 be seen to provide air passages past either side of the drives 15A-C within the cavity 40. By properly sizing the fan 110 and plenum chamber 115 to match the airflow through the cavities 40 and 45, sufficient cooling is provided to the drives and to the RAID controller board to permit long-term continuous operation. It will be appreciated that additional drives may be included in the event thinner drives are used, with
35 commensurate changes to the RAID controller discussed in connection with Figure 11.

In addition, a leaf spring 175 may be positioned at the front of the cavity 45

EXPRESS MAILING LABEL EM303714786US

9

both to urge the board 100 into the proper position and also to provide a ground plane connection to the board 100 for discharging any electrostatic charge which might be imposed on the board by the user during maintenance of the array. It will be appreciated that, unlike most subsystems within the computer system, the front 5 edge of the RAID controller 100 will be accessible to the user from the front panel of the computer system simply by removing the bezel 20. As a result, a suitable path to ground for ESD purposes is appropriate, and can be provided by plating with a conductive material a portion of at least one side of the board 100 near its front edge and connecting that plating to the case through the leaf spring 165. The leaf spring 10 is typically constructed of copper or other suitable spring material. Copper plating of such other materials may be desirable in at least some instances. The plating of the board 100 may best be seen in Figure 10A, where the plating is identified by reference numeral 285.

Further shown in Figure 4 are a pair of LEDs 180A-B for each drive, together 15 with a pushbutton 185 for each drive. The LEDs 180A typically indicate status of the associated drive and may be multicolor LEDs which use different colors to indicate different operational states. The LEDs 180B typically indicate activity on the associated drive. The pushbuttons

185 are used to signal the RAID controller that the user desires to disconnect the 20 associated drive. By depressing the pushbutton 185, the RAID controller disconnects power and signal paths to the associated drive, allowing that drive to be safely removed while the remainder of the array continues to operate. Once the drive has been electrically disconnected from the array, the drive may be physically removed by virtue of latches 65. That drive or another equivalent drive may then be added 25 back into the array by fastening latches 65. In an exemplary embodiment, insertion of the drive connector 90A into the backplane connector 90 causes the addition of the drive to be sensed by the array; however, in some embodiments the array may be caused to sense the addition of the new drive by again pushing the associated pushbutton 185.

30 With reference next to Figures 5, the rear portion of the subsystem can be seen in a top three-quarter perspective view, such that the cooling fan 110 and external connectors for connecting the subsystem to the host can be better appreciated. As discussed in connection with Figure 3, the cooling fan 110 is positioned centrally behind the backplane and spaced therefrom sufficiently to avoid 35 unacceptably turbulent airflow through fan, which increases the amount of airflow past the drives and printed circuit boards and therefore optimizes the cooling effects of the fan. At the left of the fan 110 is positioned a nine-pin D-shell connector 200,

EXPRESS MAILING LABEL EM303714785US

10

typically used to connect to a monitoring device such as the ArrayView product offered by the assignee of the present invention or other suitable device for monitoring the status of the subsystem. Below the D-shell connector 200 is a conventional power connector 205. The D-shell connector 200 and the power
5 connector 205 are, in the exemplary embodiment described herein, connected to the backplane 85 and extend through openings in the rear cover plate 105. At the right side of the fan is a high density connector 210 conforming to the single-ended Ultra-Wide SCSI standard, together with a suitable jumper block 215 for setting the ID of the unit, performing various diagnostics, and other conventional functions. The
10 connectors 210 and 215 are typically affixed to the RAID controller board 100, and extend through openings in the rear cover plate 105. The SCSI connector 210 typically provides the interface to the host system, and the entire subsystem appears as a single SCSI device to the host adapter in the host system. In other embodiments, the subsystem may comply with different interface standards such that
15 different connectors may be offered, including differential SCSI, wide SCSI, or some other interface.

Reference is next made to Figure 6, which shows the array subsystem in rear elevational view with the rear cover plate (including the fan) removed and thus shows in detail the layout of the back of the backplane 85, and Figure 7, which shows in
20 front elevational view the layout of the backplane 85. With particular reference to Figure 6, the connectors 200 and 205 can be seen to be integral with the backplane 85, and the manner by which the dual connector 95 connects the backplane 85 to the RAID controller board 100 can also be seen. In addition, a variety of vents or cutouts 225, both at the periphery and through the backplane 85, can be seen to exist in the
25 backplane to improve airflow into the plenum chamber 115. The backplane is held in place by four screws (not shown) which pass through holes 230 and mount into the mating pairs of mounting brackets 155 and 160. Also mounted on the backplane 85 is an alarm 235 which responds to signals from a variety of sensors which monitor array performance, including for example one or more drive temperature sensors
30 240, a fan sensor 245, and so on, which in the exemplary embodiment shown herein may be mounted on the front of the backplane as shown in Figure 7. In addition, the connectors 90 shown on the front of the backplane are typically high cycle, low insertion force connectors which provide both a conventional IDE bus and power to the associated drive. The drive extension 60 then provides the appropriate
35 mechanical interface to the drives, including conventional IDE connectors and conventional power connectors. Although the particular ordering of the drives 15A-C which plug into the connectors 90 is not critical, in the exemplary embodiment

EXPRESS MAILING LABEL EM303714786US

11

described herein the drive associated with the top connector is typically assigned drive 0, the middle connector drive 1, and the bottom connector drive 2.

Referring next to Figures 8 and 9, the manner by which a single drive 15A fits into a drive bracket 50 can be better understood. The drive bracket 50 comprises, as noted above in connection with Figures 1 and 2, a pair of rails 55A-B together with a drive extension board 60. The drive 15 is mechanically affixed to the bracket 50 by means of conventional machine screws, and electrically connects to the bracket through the cable 60A and the connector 90A to the backplane connector 90, as well as through a conventional Amphenol power connector 60B. The latch mechanisms 10 65 may also be appreciated.

Referring next to Figures 10¹¹~~A~~ and 11¹¹~~B~~, the layout of the RAID controller board 100 may be seen. The RAID controller board 100 comprises a single double-sided printed circuit board, the schematic of which can be better appreciated from Figure 12¹²~~A~~, discussed below. Viewed from the side shown in Figure 10¹¹~~A~~, which can be seen 15 to be the outboard side, the connectors 210 and 215 may be seen at the far left. The RAID controller includes a RAID engine integrated circuit 260 (which may be either a Field-Programmable Gate Array, an ASIC or other suitable implementation), to perform the necessary queuing and DMA functions. The RAID engine 260 communicates with cache memory 265 (Figure 11¹¹~~B~~), a RISC CPU 270 for managing 20 the operation of the RAID controller, its associated CPU memory 275 (both Figure 11¹¹~~B~~), and a SCSI processor 280 (Fig. 10¹¹~~A~~) for managing the host interface. The LEDs 180A-B and pushbuttons 185 can be seen to be connected to the RAID controller board at the forward edge (Figure 11¹¹~~B~~) while on the opposite side of the board the conductive ESD plating 285 (discussed generally in connection with Figure 25 3) may be seen. The exemplary embodiment of the RAID controller board 100 shown herein also includes a pair of connectors 275 for permitting the backplane 85 to be connected into the board 100. A time of day/date chip 290 may also be provided as well as various other sensors and logic which perform conventional functions as described in connection with Figure 12¹²~~A~~. From the arrangement of the 30 plating 285, it will be particularly appreciated from Figure 10¹¹~~A~~ that a user performing maintenance on the subsystem of the present invention is substantially prevented from damaging the RAID controller as the result of any electrostatic charge the user may carry when performing otherwise acceptable maintenance, because the plating 285 is connected directly to the ground plane as discussed above.

35 Referring next to Figures 12¹²~~A~~ and 13¹³~~A~~, the electrical operation of the invention may be better understood. In general, the RAID subsystem appears to the host system as a single volume which externally complies with conventional SCSI

EXPRESS MAILING LABEL EM303714786US

12

commands, but internally operates as a full RAID array. The RAID array operation is controlled by the RAID controller, which in turn operates by using time-division multiplexing and separate 32-bit DMA and CPU software process memory to allow for simultaneous non-contending activities at the engine's peak rate. The DMA or
5 cache memory 265, which may for example be four megabytes configured as 1x36 memory, provides a single-cycle paged EDO pipeline with bandwidth on the order of 160 MB/sec. The CPU memory 275, which may be configured in an exemplary embodiment as four megabytes configured as 1x32, provides a two-cycle paged EDO pipeline with 80 MB/sec bandwidth.

10 The CPU 270, which may for example be an LSI LR33310-40 32-bit RISC processor operating at 40 MHz, cooperates with a FLASH ROM 300 which stores an embedded RAID operating system. At the center of the architecture is the RAID integrated circuit 260, which may for example be an Altera Field Programmable Gate
15 Array or equivalent or may be configured as an ASIC, which provides command queues for each DMA I/O channel, manages the various data I/O queues, manages the bus activities on the key buses associated with it, and supports system peripheral functions. Five major buses are associated with the FPGA 260: a 40 MB/sec, 16-bit SCSI processor bus 305 (typically configured for Ultra-SCSI operation although other SCSI protocols can be supported); a 3.33 MB/sec 8-bit SCSI chip pipelined I/O bus
20 310; a 16 MB/sec 16 bit IDE drive bus 315; a 160 MB/sec 36-bit disk cache memory (DCM) bus 320; and an 80 MB/sec 32-bit CPU bus 325. The FPGA 260 is configured to permit operation of all five buses in parallel, with the RAID engine 260 operating at a sufficient speed to multiplex the access to the RAID engine 260 by the SCSI processor bus 305, the IDE bus 315, and the DCM bus 320 by allocating, within
25 a defined cycle, one time slot for each of the SCSI processor bus 305 and the IDE bus 315, and two time slots for the DCM bus 320. In the exemplary embodiment discussed herein, a complete cycle may be on the order of 100 ns, with each of the four time slots allotted 25 ns. Because of the performance of the RAID engine 260, the net subsystem throughput is primarily dependent on four factors: the
30 performance of the IDE drives, the RAID function overhead in the embedded operating system, the performance of the user's host adapter, and the driver overhead of the user's host application.

Still referring to Figure 4¹², the operation of the system is substantially as follows: On power-up, the system comes to a stable state by loading the operating
35 system from the Flash ROM 300 into CPU memory 275 associated with the RISC processor 270 and initializing the remainder of the system to known states. At some point after initialization, a request either to read or to write will be received from the

EXPRESS MAILING LABEL EM303714786US

13

host system 340 at the host SCSI bus 350, which may be terminated by a termination block 355 if appropriate. The request is then handled by the SCSI processor 280, which sends the appropriate signals to the RAID Engine 260 over a pipelined I/O bus 310 and receives back the appropriate confirmation signal. The data is then made
 5 available by the host system over the SCSI DMA bus 305. At this point the Disk Cache Memory 265 is empty. If the request is to write information, the CPU 270 instructs the RAID engine 260 to pass the data to the DCM 265, where it can be maintained in cache. Thereafter, during background processing, the data can be written to assigned disk(s) as appropriate (in accordance with the RAID striping being
 10 used) by first having the data accessed by the RAID engine 260 over the bus 320 and written out over the IDE bus 315 to ISO disk buffers 360A-C. The data is then written to the specific disks 15A-C. It will be appreciated that the bus 320 comprises, in the exemplary system described herein, ten address lines and 36 data lines. Likewise, the bus 325 comprises thirty-two address lines and thirty-two data lines.
 15 The process ends with a confirmation signal supplied from the RAID engine to the SCSI processor 280 and thence to the host 340. The timing of the various events will be discussed in greater detail in connection with Figures ^{13 to 16} ~~12 and 13~~ ¹⁷ ~~14~~. In figure 13, DP2 and DP0 represent "data path in" and "data path out", respectively.
 In a read operation, the process is substantially similar, though somewhat reversed. The process begins by enabling the SCSI interface to be active, typically
 20 done at startup. The host then sends a confirmation/acknowledge signal and executes a set-up, followed by sending a request for specified data over the PIO bus 310. The request is then detected by RAID engine 260, which passes it to the drives 15. The data is returned from the drives to the RAID engine 260, where it is passed to the disk cache memory 265 for interim storage. At the appropriate time, the CPU
 25 270 causes the data to be read from the DCM 265 via the bus 320 and passed through the engine 260 to the SCSI processor 280 over the data bus 305. The data is then passed from the SCSI processor 280 to the host over the bus 350.

In addition to its data handling functions, the RAID engine also manages a number of peripheral housekeeping functions. Included among these are monitoring
 30 of the over-temperature detector 240 and the fan error detector 245, generating alarm signals (when appropriate) at the alarm 235. The time-of-day/date clock 290 is also monitored, for which power is supplied by a battery or other power source 365 when the system is off. Hardware sense lines 370 can be monitored by means of status registers 375. Power surge control for the drives can be monitored at buffer
 35 380. Monitoring of the subsystem is also provided over a duart 385. Typical monitoring is performed over RS232 links 390 for both monitoring and maintenance.

Referring next to Figures ^{13 to 16} ~~12 and 13~~ ¹⁷ ~~14~~, the details of the operation of the RAID

EXPRESS MAILING LABEL EM303714786US

14

engine 260 can be better appreciated, including the timing by which the signals on the various buses are multiplexed by the RAID engine. As with the prior figures, like elements have been assigned like reference numerals from Figure 14.

As before, when the host system is turned on, the RAID subsystem 10
 5 initializes and the RISC CPU 270 generates a series of enabling factors as established by the software maintained in the FLASH ROM 300. The enabling factors place the IDE drives in known states and also place the SCSI processor 280 in an active and enabled state, including notifying the host system 340. The host system confirms and acknowledges the notification from the SCSI processor. In
 10 addition, the enabling factors place the RAID engine 260 in a known state, and in particular initialize a 40 MHz I/O Command Queue Processor 390 which is internal to the RAID engine 260. *Figure 15 shows an alternative arrangement of a portion within a dot-and-dashed line in Figure 14.*

After initialization, the host system sends data to be written to the drives 15A-C, as before. The information, which comprises header information and data, is
 15 supplied to the SCSI processor 280 over the bus 350. After processing by the SCSI processor 280, the header information is supplied to the RAID engine 260, indicated by the dashed line in Figure 14, over the eight-bit programmable I/O bus 310. The header information is supplied to the RAID engine 260 through a SCSI-PIO 400, which has a data path in 405 and a data path out 410. The data path in 400 links to
 20 a one side of a mux 415, which in turn feeds, indirectly, the input to the I/O command queue processor 390. The I/O command queue processor 390 is a frame-based script processor and supplies half-word commands, row address commands and column address commands to a register 395 via a 16-bit bus. The register 395 can also receive addresses from the RISC processor 270; the processor 270 can also
 25 supply addresses to the DCM 265 via a 10-bit branch 397 of the bus 325, fed through a mux buffer 399. The output of the register 395 can be supplied via a multiplexed 10-bit bus (addressing one megabyte of address space) to the DCM 265 through a pipeline register 401 and buffer 403. The output of the register 395 provides, indirectly, the data-path-out referred to elsewhere in connection with the RAID engine
 30 260, including the second input to the mux 415.

Concurrently with the header information supplied from the SCSI processor 280 to the SCSI PIO 400 on the bus 310, the data from the SCSI processor 280 is supplied to the RAID engine 260 via a sixteen-bit DMA bus 305. In particular, the data is fed to a 16-bit-to-32-bit funnel 420, operating at 40 MHz, because the RAID
 35 engine 260 operates internally at 32-bit width. The data is supplied to one side of a funnel mux 425 and then to one side of a I/O queue mux 430. The output of the queue mux 430 is supplied to a frame-based I/O queue 435, operating at 40 MHz.

EXPRESS MAILING LABEL EM303714786US

15

and configured at 256 x 32 to provide 160 MB/sec throughput. Other inputs to the I/O queue 435 include various IDE pointers 440, SCSI pointers 445 and DCM pointers 450. The data is clocked through to the output of the I/O queue 435 and supplied to a first pipelined output register 455, and then to one side of a DCM mux 5 460. The output of the mux 460 is provided to a second pipelined output register 465, through a buffer 470 and then out of the RAID engine 260 to the DCM 265.

The data is stored in the DCM 265 until appropriate for writing to appropriate ones of the disks 15A-C, typically determined by the RAID operating system according to conventional algorithms. At that time, the I/O command queue 390 10 issues a command to write the data to the disk drives. The data is supplied by the DCM 265 to a buffer 475 and then to a pipelined input register 480. The data is then provided to a second input register 485 as well as one side of a processor input mux 490. To write to the drives, the data is fed through the register 485 to the other side of the mux 430, and then to the I/O queue 435.

15 The data out of the I/O queue 435 is provided to the SCSI I/O funnel 420, but is also provided to a disk I/O funnel 500. The disk I/O funnel 500 reconverts outgoing data from a 32-bit data width to a 16-bit data width for communication with the disk drives 15. The remainder of the communication to the disk drives is as described in connection with Figure 12.

20 Retrieving data from the RAID subsystem is the other operation typically required of the RAID subsystem 10 by the host system 340. Retrieving data is initiated from the host system 340, which again supplies the host's request to the RAID engine 260 via the PIO bus 310, the SCSI PIO 400, and then the data-path-in 405 to the I/O command processor 390 through the mux 415. The I/O command 25 processor 390 then supplies the appropriate RAC/CAC addresses via the register 395 to cause the data to be retrieved.

The appropriate addresses for the data desired by the host system are supplied to the DCM 265. If the data is maintained in the DCM 265, it is supplied via the registers 480 and 485 to the mux 430 and then to the I/O queue 435. From the 30 I/O queue 435 the data is supplied to the SCSI I/O funnel 420 where the outgoing data is converted to 16-bit width. The data is then supplied on the DMA bus 305 to the SCSI processor 280, and finally out to the host 340 over the bus 350.

However, if the data requested by the host is not currently maintained in the cache 265, the data must be requested from the disks. In this instance, the 35 addresses for the requested data are supplied via the registers 480, 485 and mux 430 to the I/O queue 435. The output of the I/O queue 435 is then supplied to the disk I/O funnel 500 and out to the drives 15A-C. The data is then retrieved from the

EXPRESS MAILING LABEL EM303714786US

16

drives after the required latency, after which the data incoming from the drive is converted from 16-bit width to 32-bit width in the disk funnel 500. The output of the funnel 500 data is then supplied to the second side of the mux 425, and from there to the I/O queue 435 through the second mux 430.

5 The output of the I/O queue 435 is then fed through the I/O funnel 420 in the same manner as described above for outgoing data, with the result that the data is supplied to the host system in the conventional manner. For implementations based on a field programmable gate array (FPGA), an FPGA boot ROM 492 may be provided to personalize the FPGA upon power-up. In ASIC or other gate array
10 implementations, such a boot ROM is not necessary. Likewise, the RAID OS is loaded into CPU memory 275 upon power-up, and all software control is derived from the instructions stored in the CPU memory.

A key feature of the operation of the present invention is that a single RAID engine 260 is able to manage the multiple DMA queues necessary to communicate
15 with the SCSI processor, the disks, and the processor 270. This objective is achieved by time multiplexing the key buses which provide data to the RAID engine 260. This is possible because the I/O queue operates at an effective rate of 160 MB/sec, compared to the other devices which operate at no more than 40 MB/sec. This allows the RAID engine to allocate approximately one fourth of its time to each
20 of the SCSI processor and disk drives, and to allocate approximately one-half its time to DMA addressing. Shown as Figure ~~14~~¹⁷ is a timing diagram which provides the phased accessing necessary to provide the time division multiplexing important to some aspects of a presently preferred embodiment of the invention.

In particular, the 40 MHz clock of the I/O queue 435 and script processor 390
25 is shown at 600, while the SCSI Phase for access to the RAID engine is shown at 605. The IDE phase is shown at 610, while the DCM phase is shown at 615. In an additional feature, in the event that a cycle occurs when no access is requested from the IDE drives, the phase is reassigned for use by the DCM. Similarly, for those cycles in which the SCSI processor requires no I/O access the phase allocated for
30 the SCSI processor is reassigned to the DCM. It will thus be appreciated that extremely high throughputs can be achieved with the present design.

Referring next to Figure ~~14~~¹⁸, the hot-swapping arrangement of the present invention -- by which one or more drives may be removed while the remainder of the array continues to operate -- may be better understood. In particular, in the event
35 the user desires to remove one of the drives 15A-C, for example due to the failure of a drive, the user actuates pushbutton switch 185 associated with the drive to be removed. This signals the CPU 270, which operates under software control to signal

EXPRESS MAILING LABEL EM303714786US

17

the FPGA control logic to power down both the 12 volt and 5 volt supplies 700 and 705, respectively, of the associated drive. In addition, the data path 710 and control path 715 are caused by the processor to be electrically disconnected from the remainder of the subsystem. At this point, the user can readily undo the latches 65
5 and remove the necessary drive.

To reinstall the drive, the user simply reverses the mechanical portion of the process by inserting the drive into the drive bay and latching the latches 65. A bistable latch senses the reinsertion of the drive, and signals the CPU 270 to reapply both power and signal connections to the newly-installed drive. In this manner the
10 old drive may be removed and the new drive installed.

It can therefore be appreciated that a new and novel system for a RAID array subsystem and highly integrated controller has been described. It will be appreciated by those skilled in the art that, given the teachings herein, numerous alternatives and equivalents will be seen to exist which incorporate the invention disclosed hereby.
15 As a result, the invention is not to be limited by the foregoing exemplary embodiments, but only by the following claims.

EXPRESS MAILING LABEL EM303714786US

18

WE CLAIM:

1. A mass storage array subsystem comprising

a first longitudinal cavity for housing a mass storage controller which includes a SCSI interface to a host system and an IDE interface to internal mass storage devices,

a second cavity for housing a plurality of IDE drives capable of communicating with the host system such that the longitudinal axis of the first cavity is parallel to the longitudinal axis of the second cavity,

a backplane having vents therethrough and adapted to connect to each of the IDE drives housed within the second cavity, and

a plenum chamber formed by the backplane and the case to provide efficient airflow through the first and second cavities.

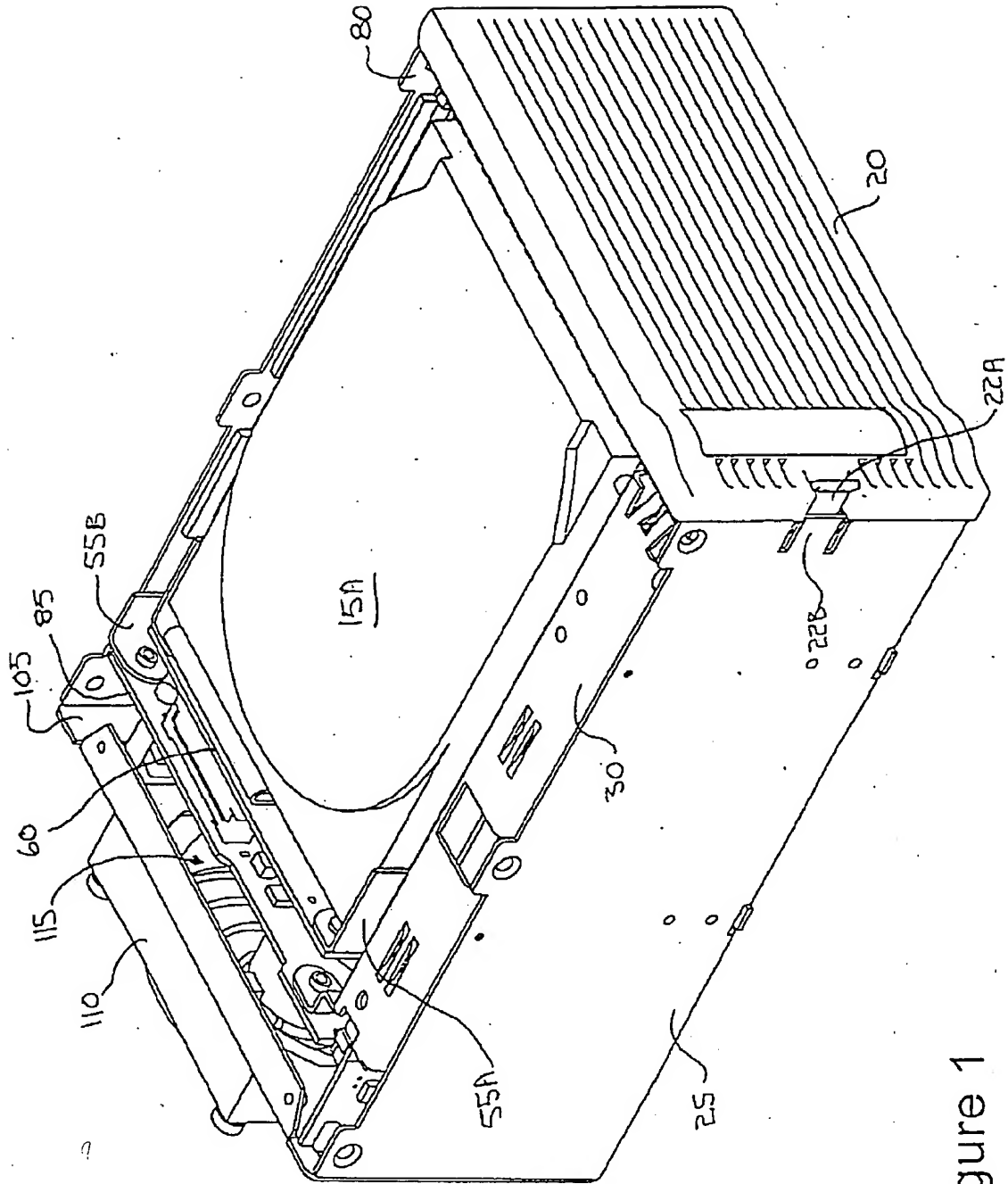
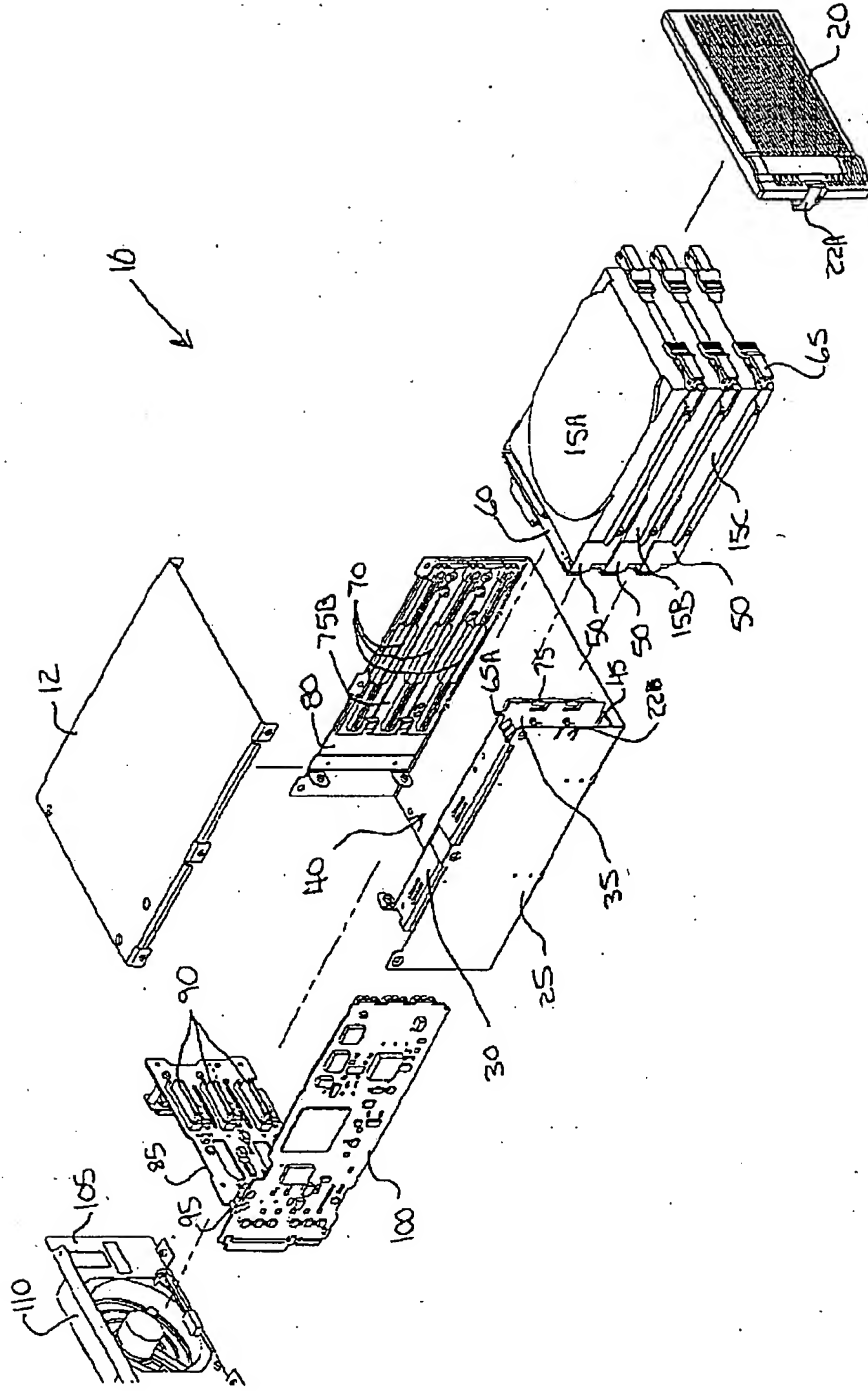


Figure 1



SCALE 0.500 SCALE 0.500 SCALE 0.500 SCALE 0.500 SCALE 0.500 SCALE 0.500

Figure 2

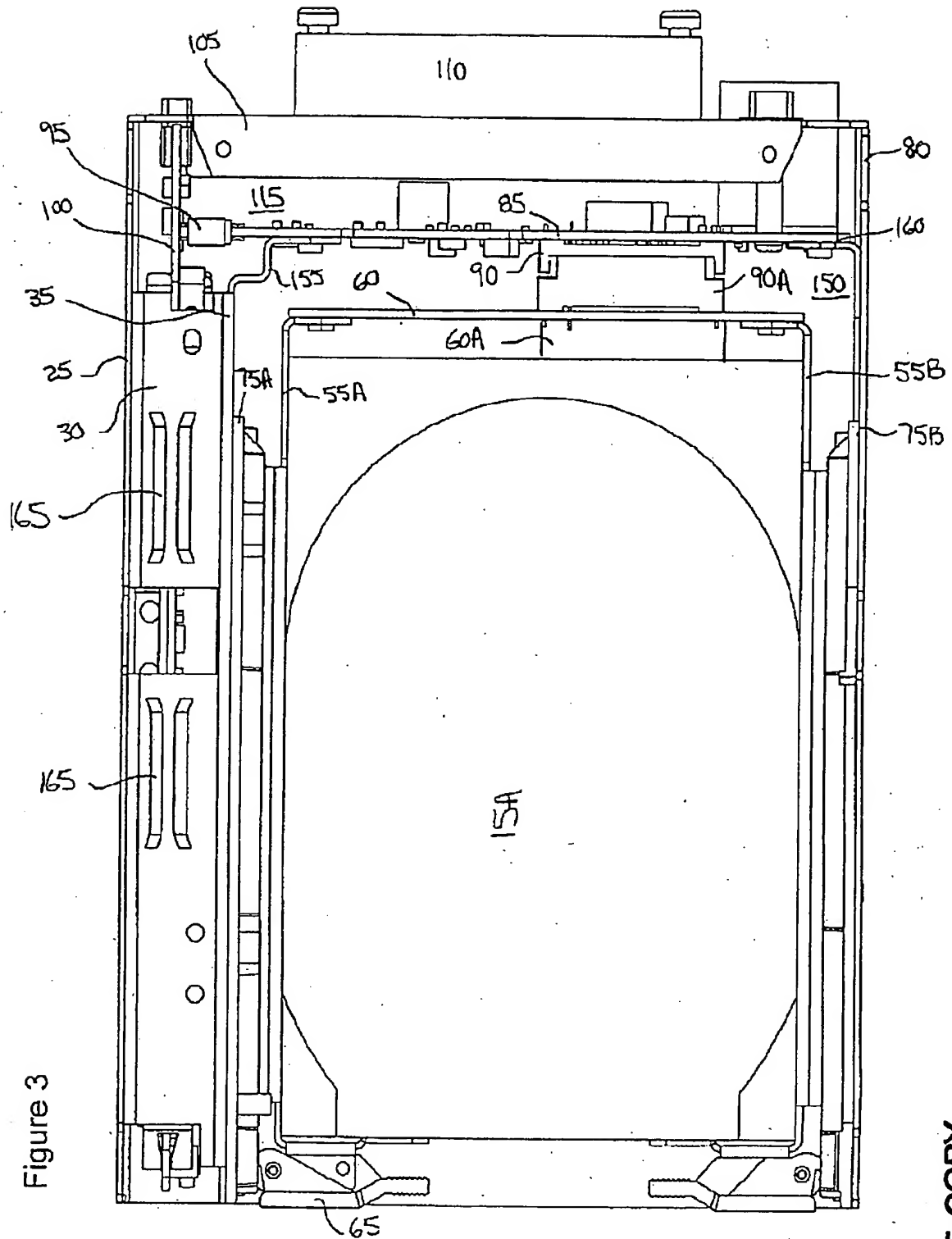


Figure 3

BEST AVAILABLE COPY

Figure 4

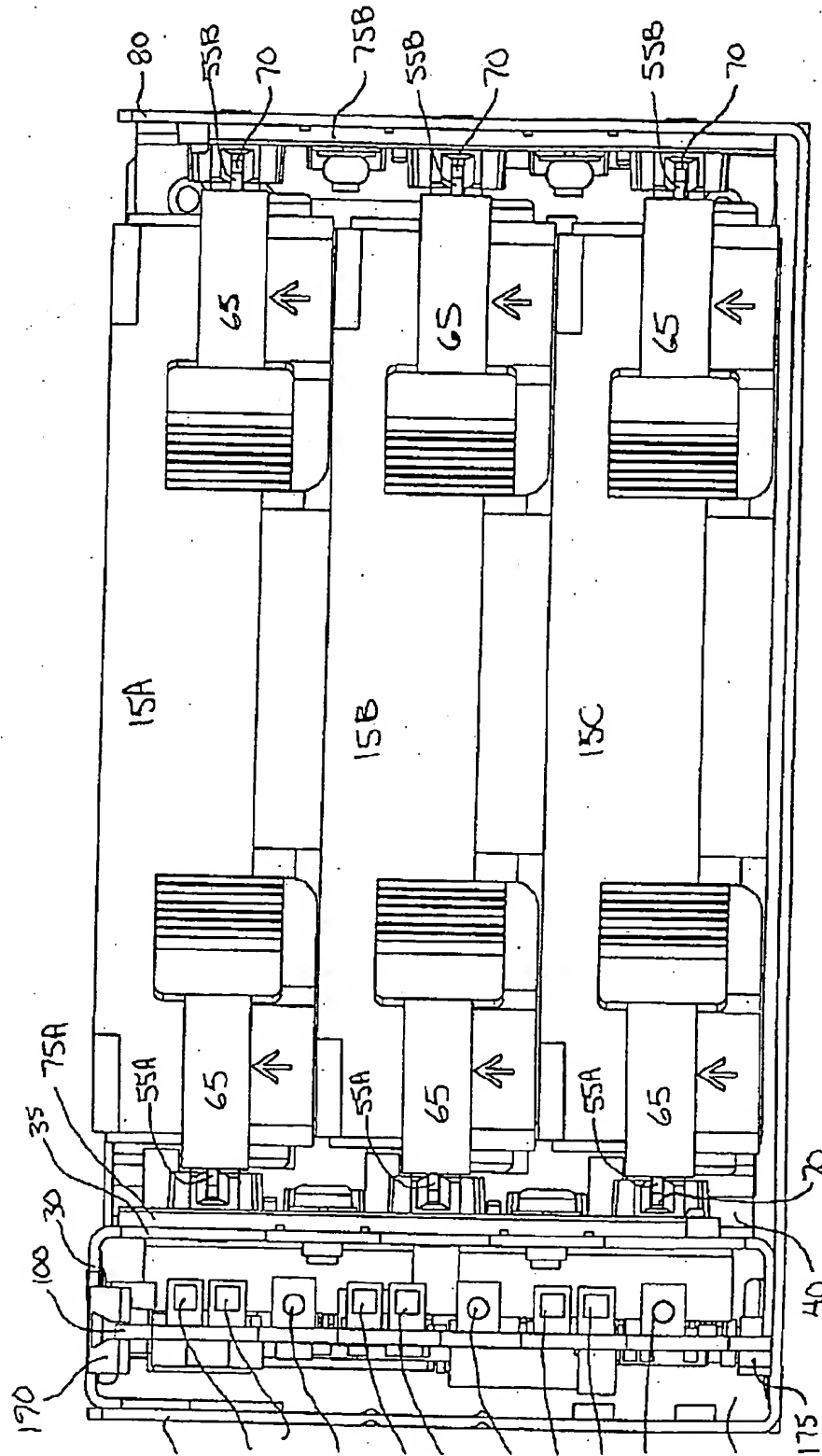
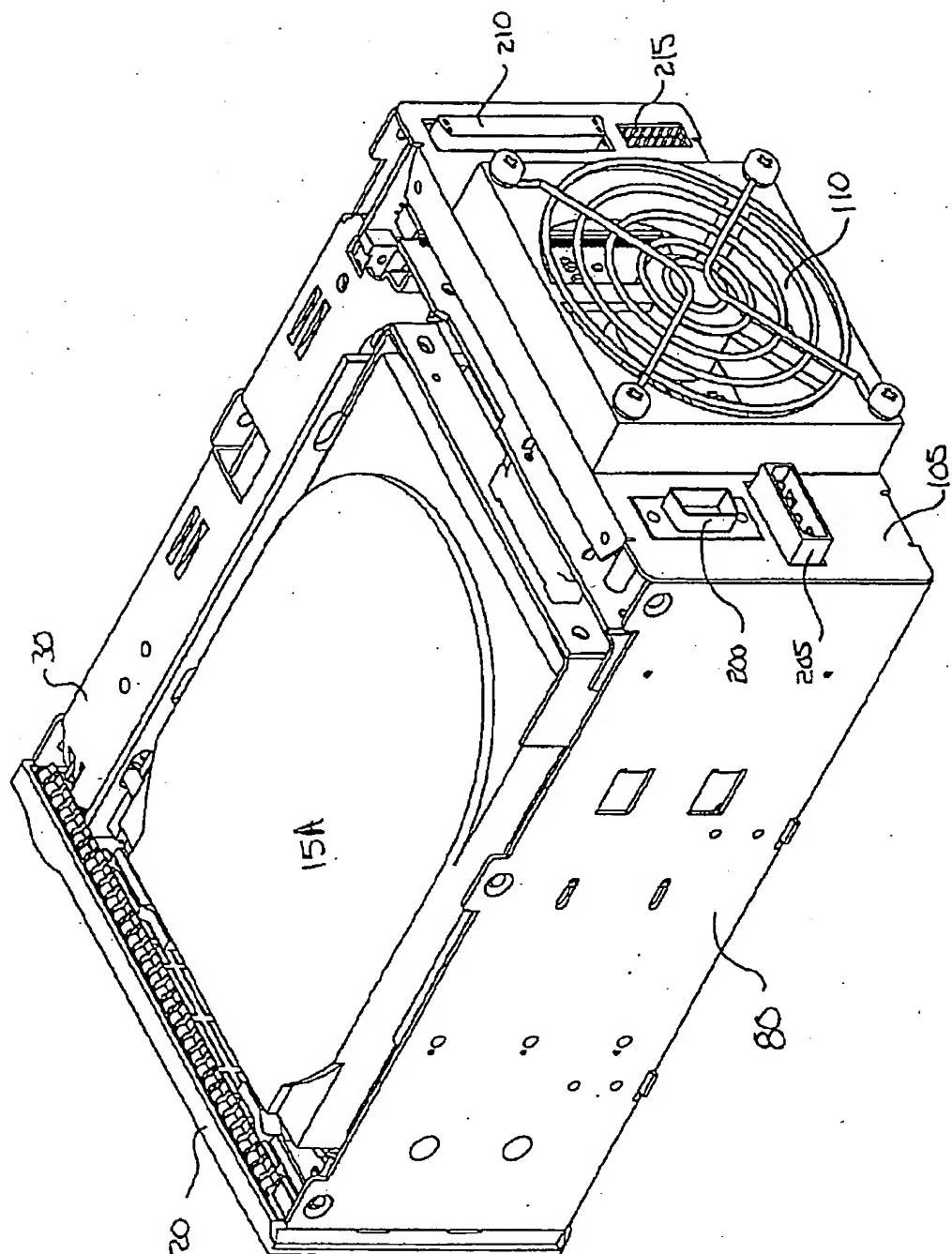


Figure 5



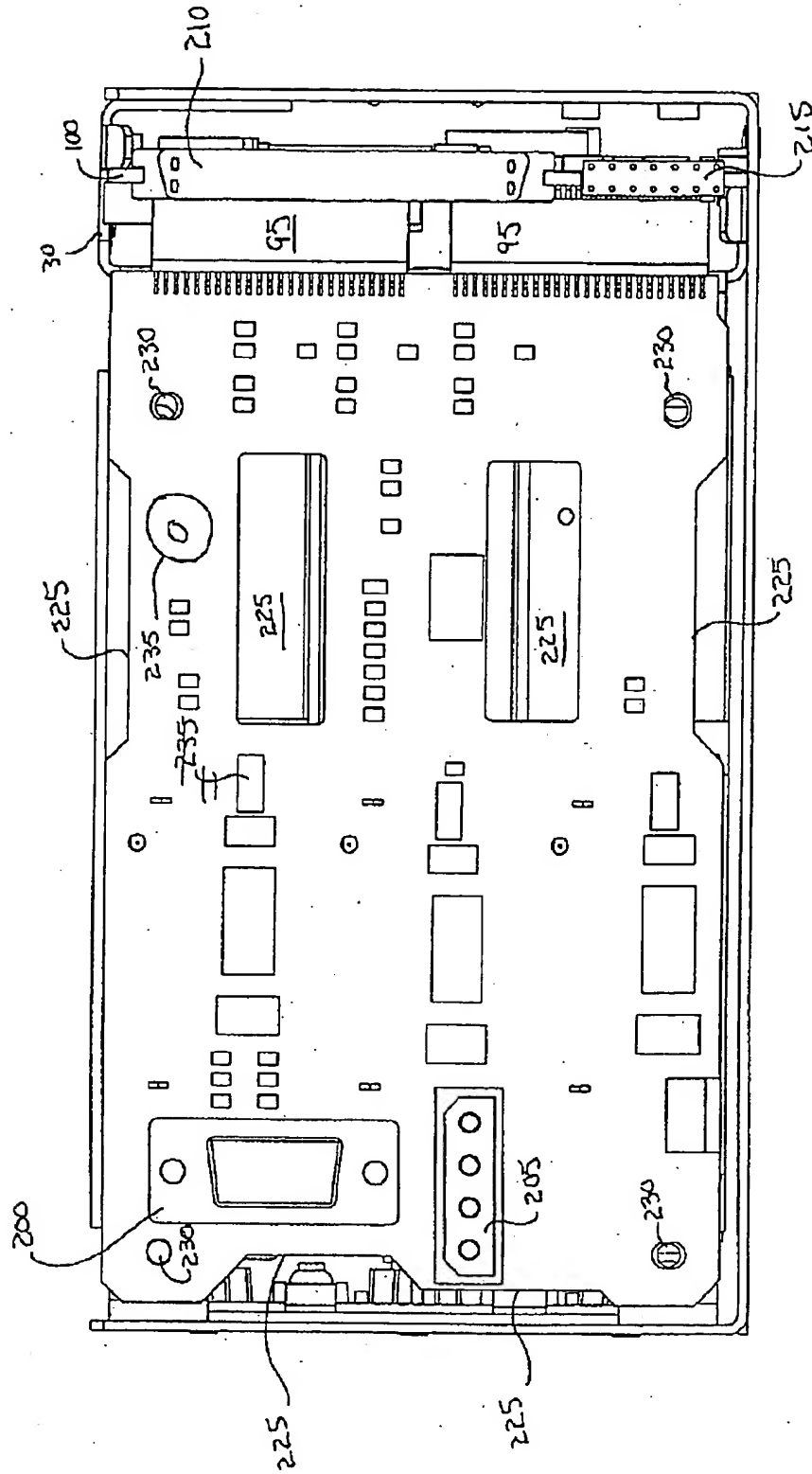


Figure 6

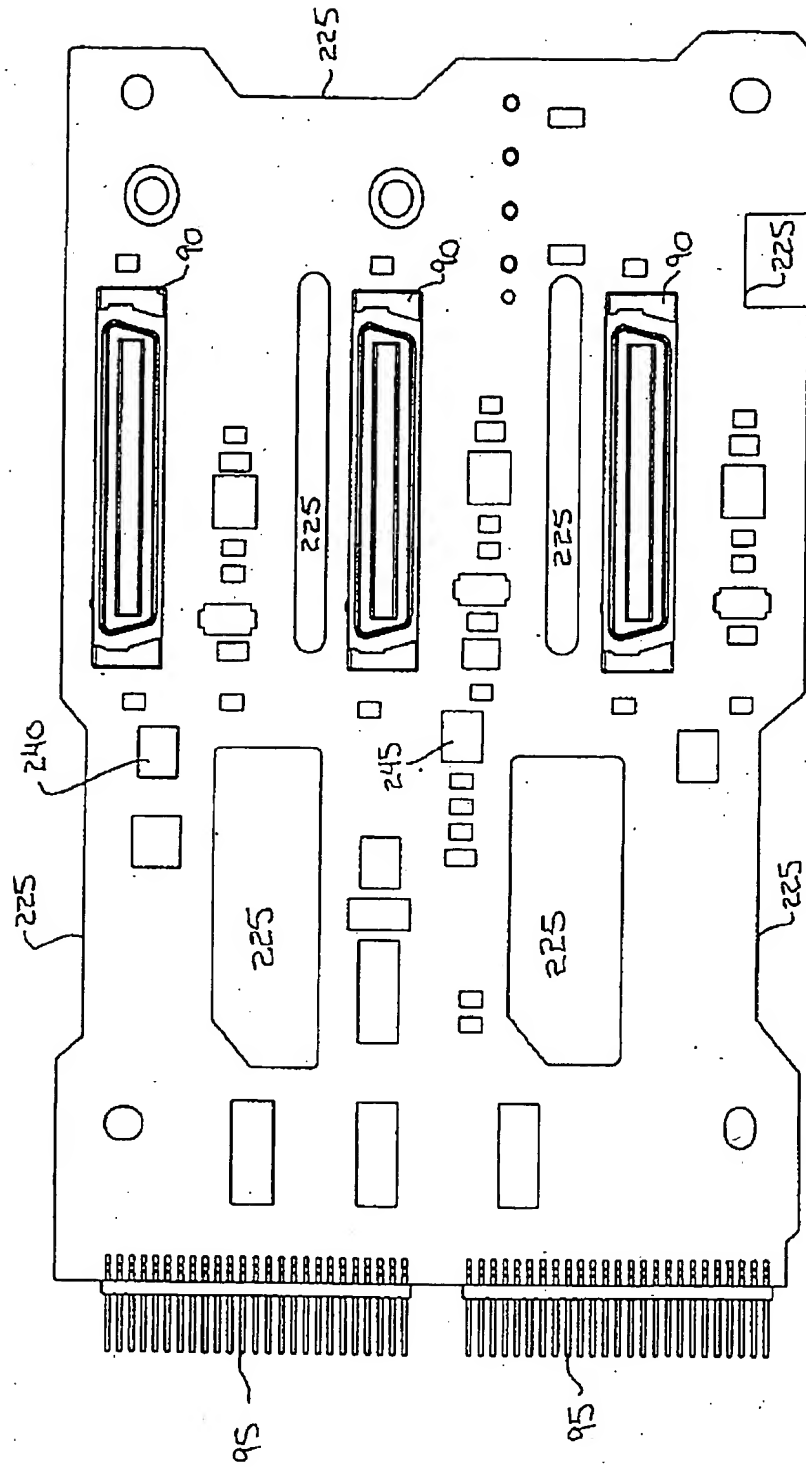


Figure 7

Figure 8

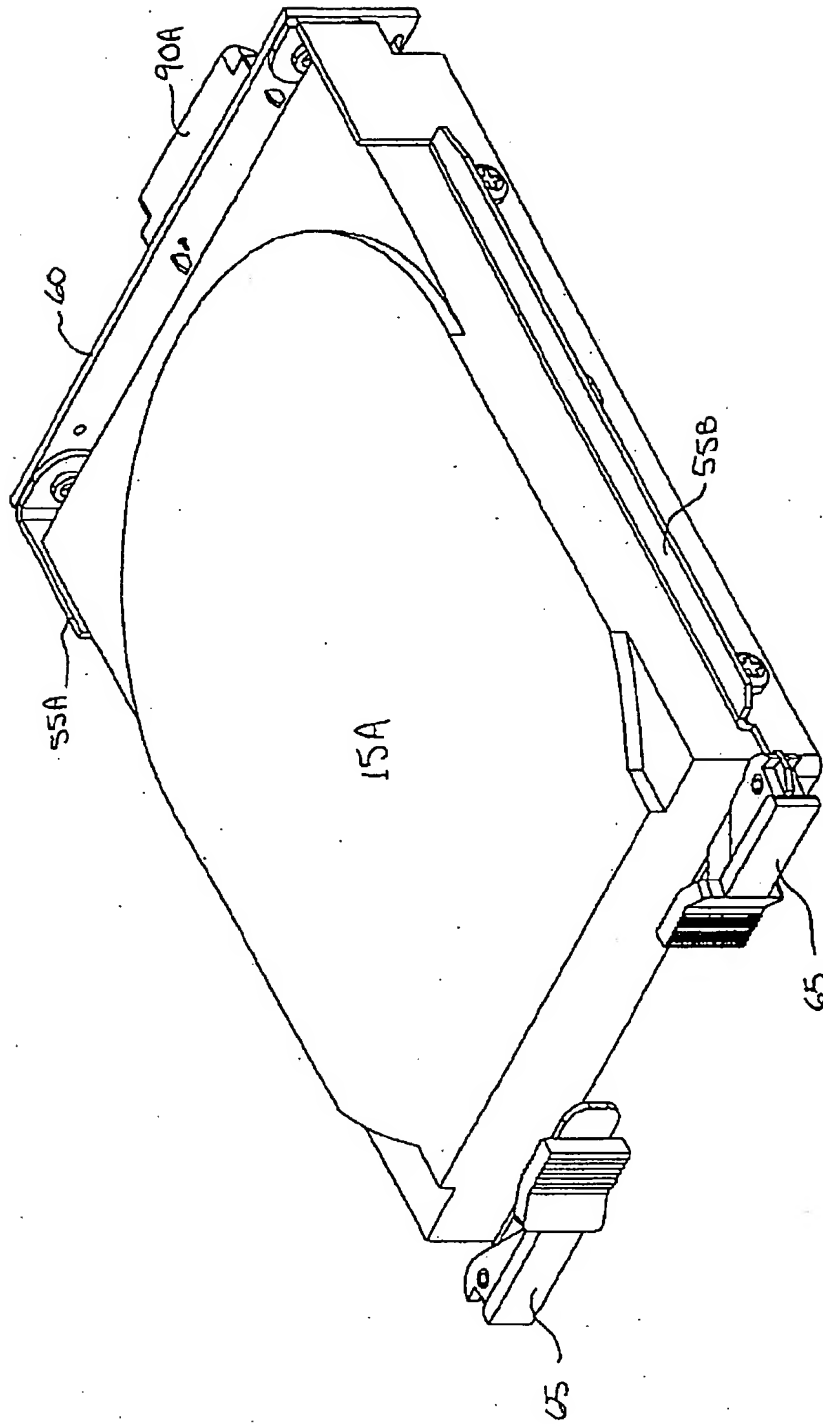
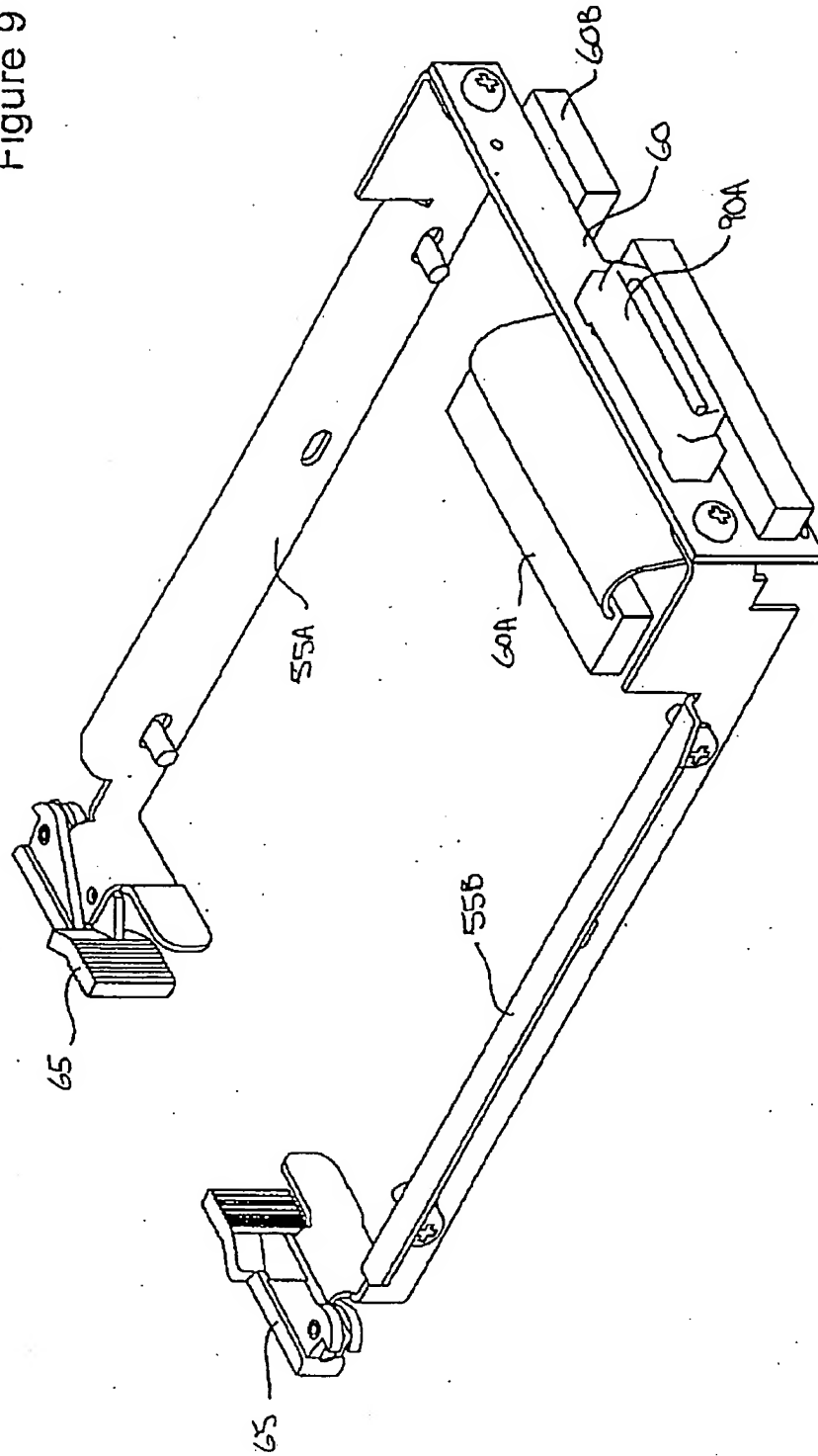


Figure 9



BEST AVAILABLE COPY

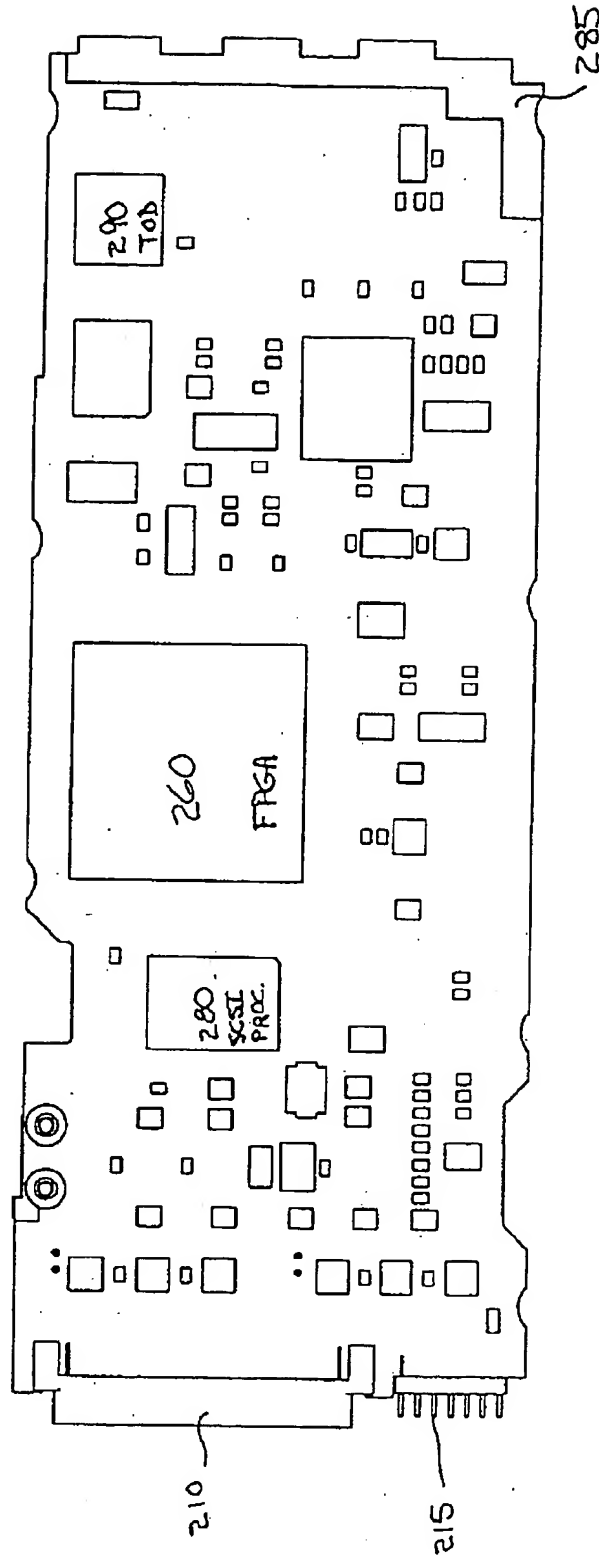


Figure 10A

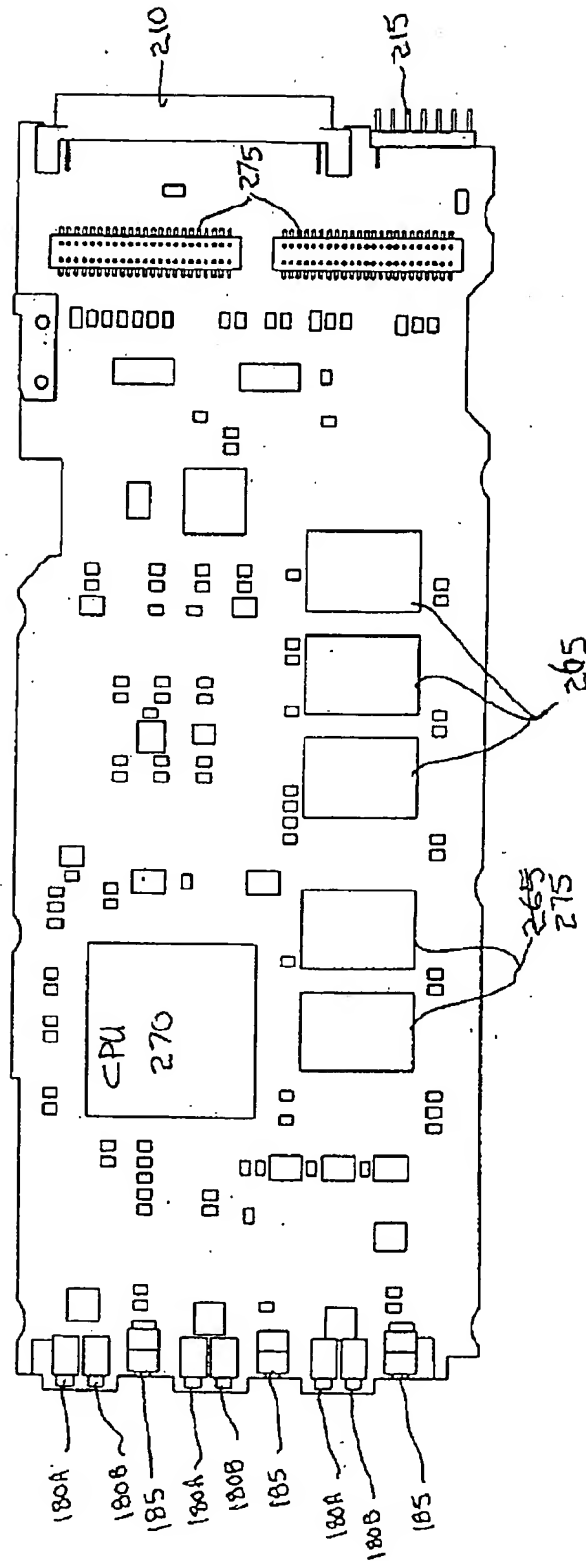


Figure 10B

BEST AVAILABLE COPY

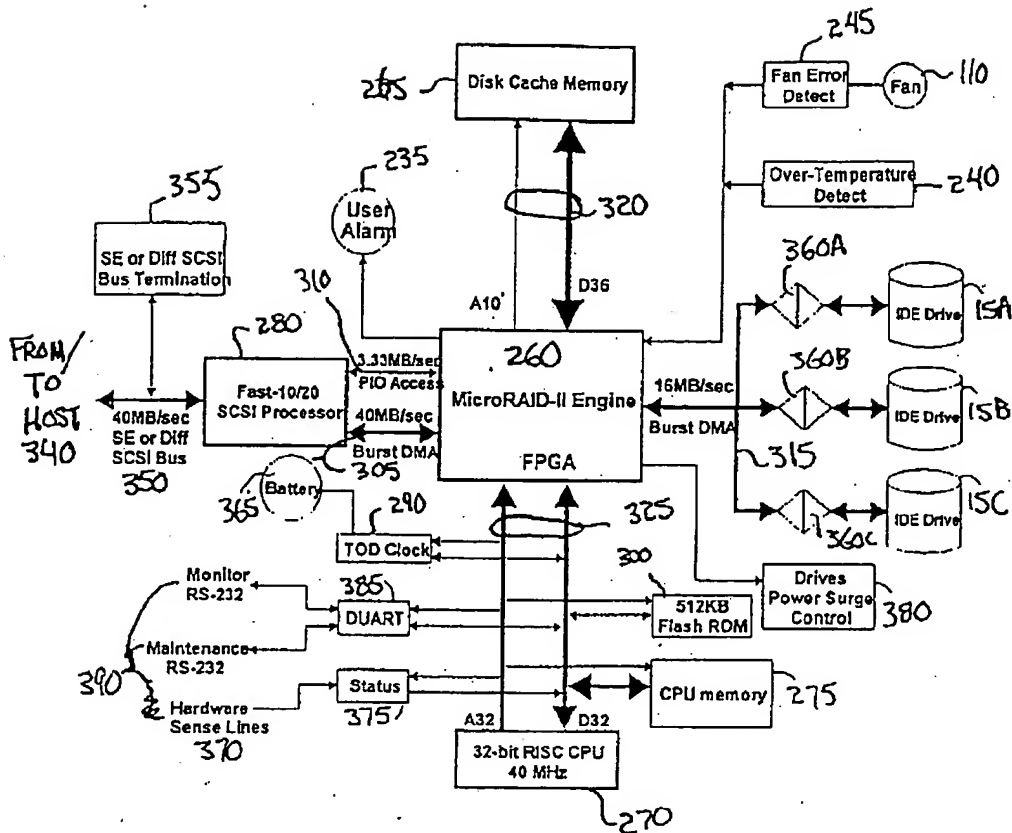


Figure 12

BEST AVAILABLE COPY

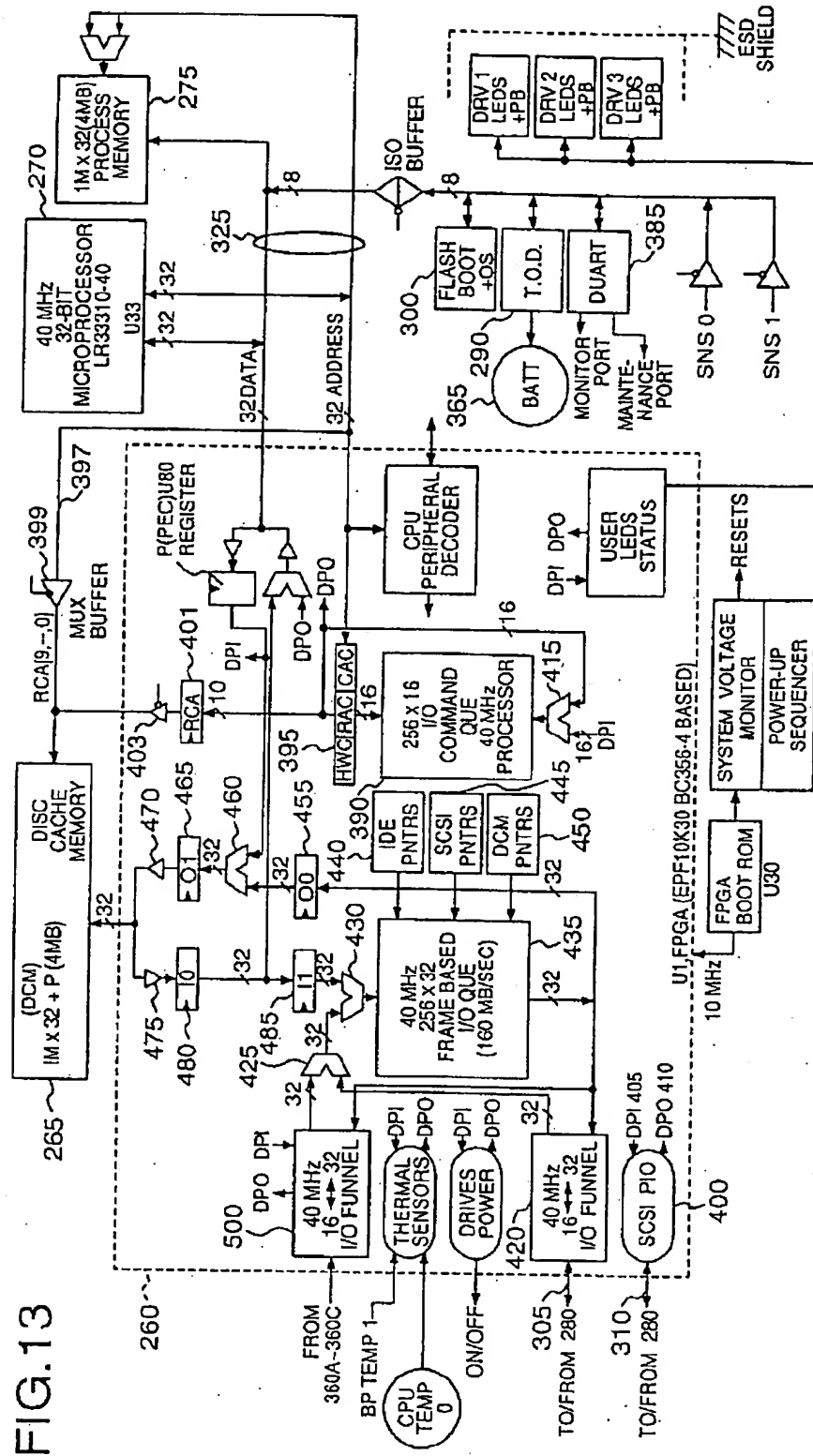


FIG.14

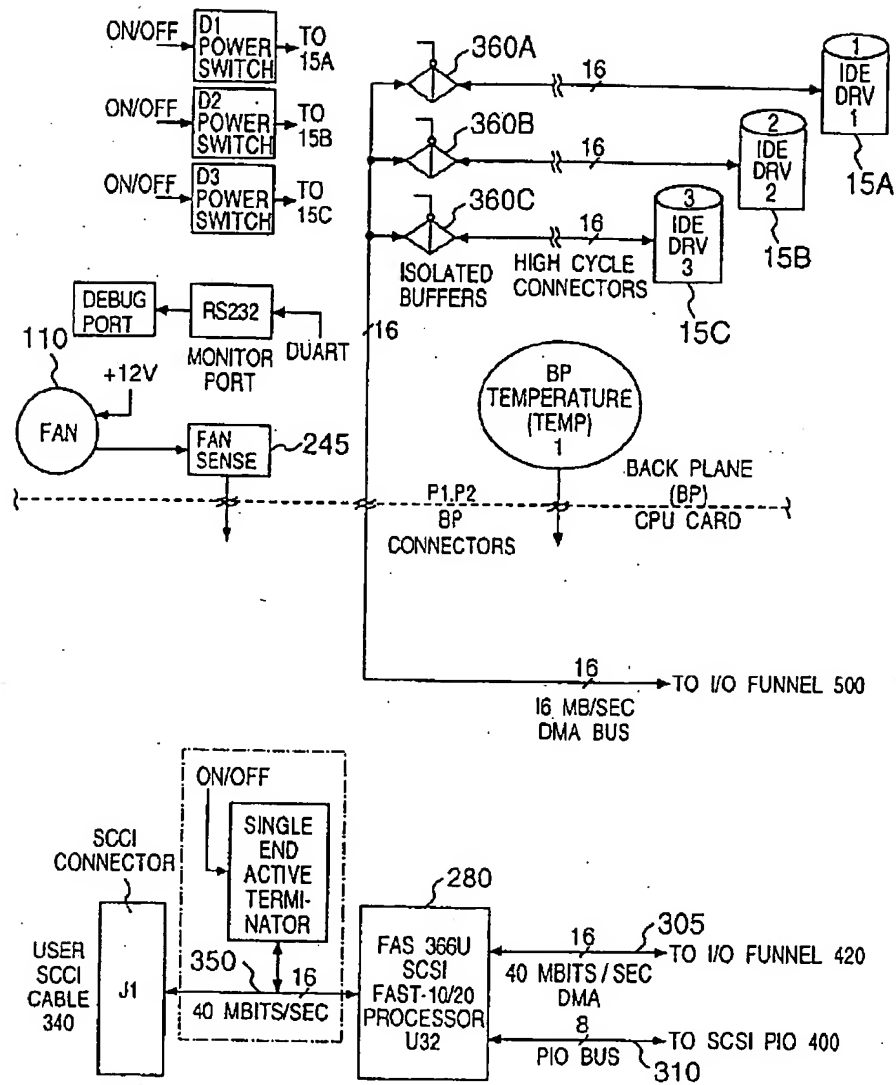


FIG.15

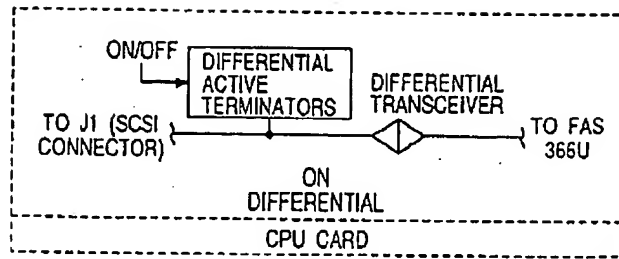
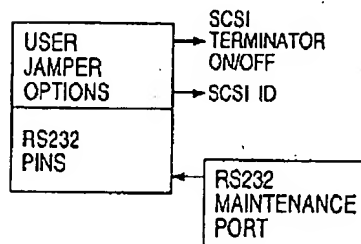


FIG.16



BEST AVAILABLE COPY

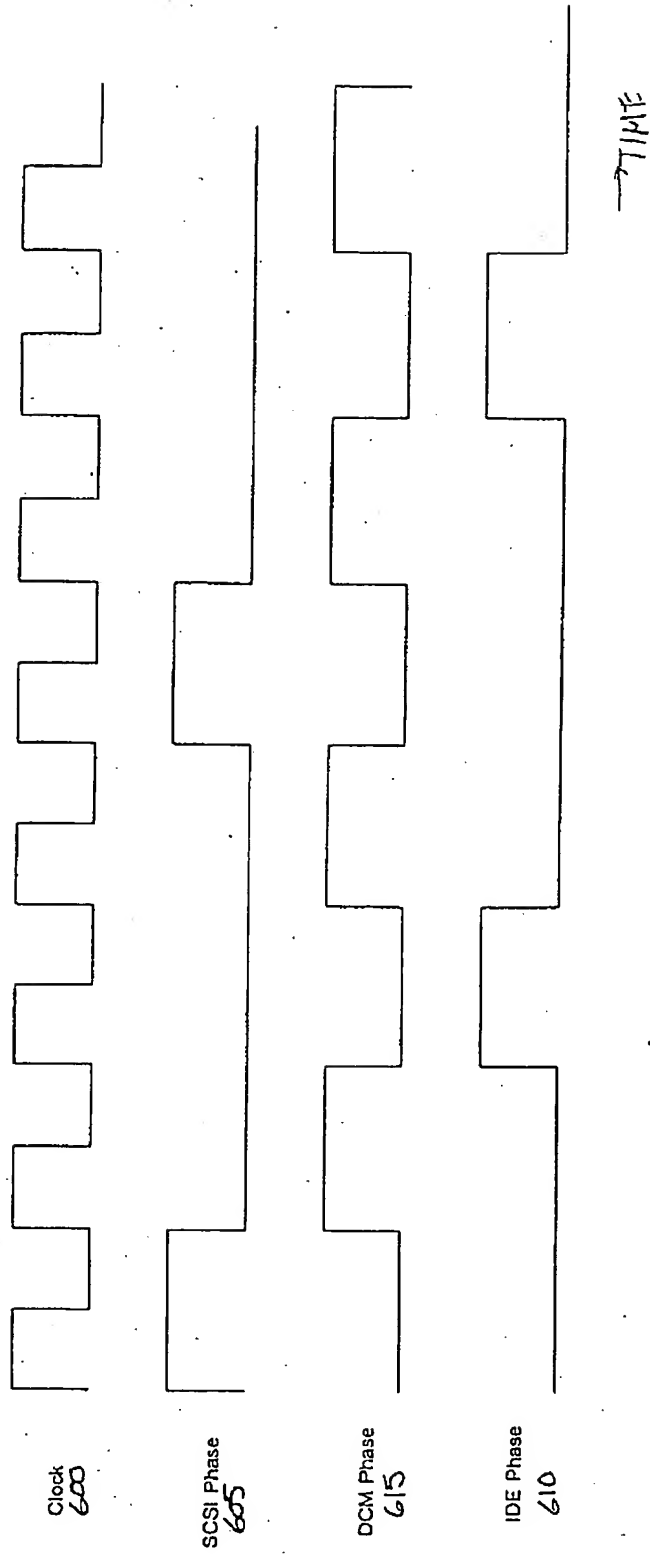


Figure 17

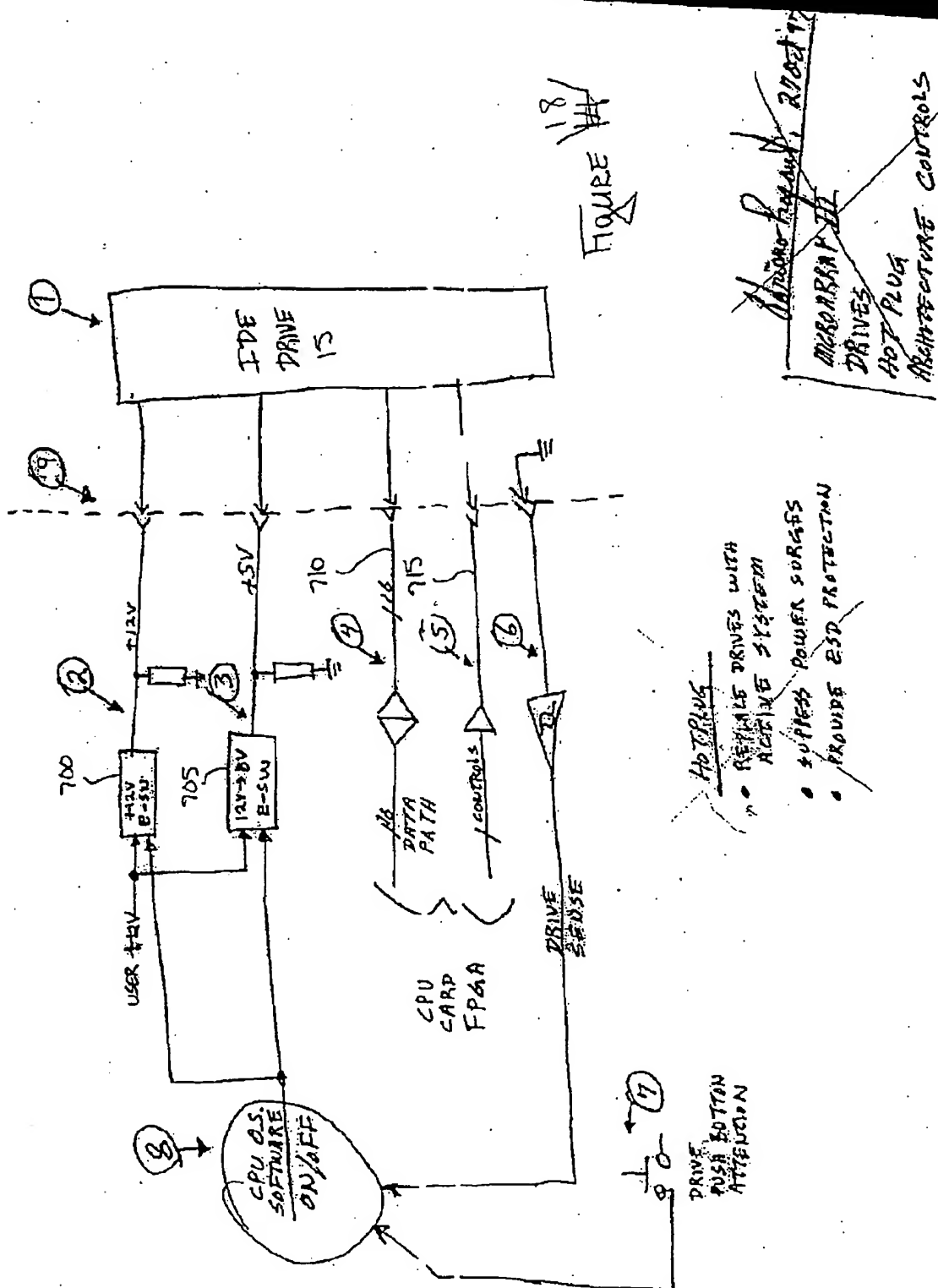


Figure 18

- NO TAPPING
- REPAIRABLE DRIVES WITH ACTIVE SYSTEM
- SUPPLIES POWER SURGES
- PROVIDES ESD PROTECTION

~~HARDWARE~~

BEST AVAILABLE COPY

1 Abstract

ABSTRACT OF THE INVENTION

The present invention provides a method and apparatus for a mass storage subsystem such as a RAID array. The invention includes a housing which defines first and second cavities with the first cavity housing an array controller such as a RAID controller. The second cavity houses a plurality of substantially conventional IDE drives conforming to the 3.5" form factor. The array is configured to maximize cooling of the array controller and the drives within the extremely small space defined by the housing.

2 Representative Drawing

Fig. 2